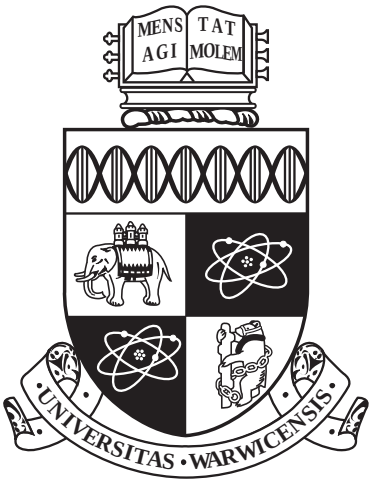


Analysis of Data assimilation schemes

by

Abhishek Shukla

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Mathematics

September 2016

THE UNIVERSITY OF
WARWICK

Contents

Acknowledgments

It is a pleasure to thank the many people who made this thesis possible. First and foremost I would like to thank my supervisor Andrew Stuart for his constant support, patience and guidance. I am grateful to him for introducing me to the field of data assimilation and giving me the opportunity to work on this thesis. On a more personal level I owe my deepest gratitude for the endless support he provided during some of the most difficult times. I also want to thank Kody Law for hours of discussions, for his supervision and friendship.

Many other people from the data assimilation community have been of great help to achieve this work. I am heartily thankful to them for the nice discussions, in person or by e-mail, for their advice and the support: Henry Abarbanel, Amit Apte, and David Kelly. I am also indebted to the useful discussions with Gareth Roberts and Jon Warren which have always been of great help. I am endlessly grateful to Andreas Dedner, Bjorn Stinner and Charlie Elliott for their support in everything.

I would like to thank my MASDOC cohort for their support and encouragements. I would also like to thank Daniel Sanz-Alonso for his contributions to this work. Lastly, and most importantly, I am deeply indebted to my parents for their endless love, for all the sacrifices they made and for supporting me all the way long even in the hardest times. To these wonderful people, I dedicate this thesis.

Declarations

This thesis is divided into 5 chapters. The first chapter is a brief introduction to the data assimilation schemes used in this thesis. This chapter also serves the purpose of establishing the notation for the rest of the thesis and does not include any original contribution.

Chapters 2 and 3 consist of published papers and present the research I have performed for my Ph.D. Chapter 2 is a transcript of the paper [40] co-authored by my supervisor Andrew Stuart and collaborator Kody Law. My contribution to this paper was in proving theoretical results and performing initial numerical experiments.

Chapter 3 is the transcript of the paper [39] written in collaboration with Andrew, Kody and Daniel Sanz-Alonso. My contribution to this work was in proving theoretical results and performing numerical experiments. Proof of the Theorem 3.4.6 was provided by Daniel. I wrote the first draft of both the papers which were further edited and improved upon by my co-authors.

Chapter 4 and 5 are based on the discussions with Henry Abarbanel and Andrew Stuart on the implementation of 4DVAR .

This thesis has not been submitted for a degree at any other university. It has not been submitted for award at any other institution for any other qualification.

Abstract

Data assimilation schemes are methods to estimate true underlying state of the physical systems of interest by combining the theoretical knowledge about the underlying system with available observations of the state. However, in most of the physical systems the observations often are noisy and only partially available. In the first part of this thesis we study the case of sequential data assimilation scheme, when the underlying system is nonlinear chaotic and the observations are partial and noisy. We produce a rigorous and quantitative analysis of data assimilation process for fixed observation modes. We also introduce a novel method of dynamically rearranging observation modes, leading to the requirement of fewer observation modes while maintaining the accuracy of the data assimilation process.

In the second part of the thesis we focus on 4DVAR data assimilation scheme which is a variational method. 4DVAR data assimilation is a method that solves a variational problem; given a set of observations and a numerical model for the underlying physical system together with a priori information on the initial condition to estimate the initial condition for the underlying model. We propose a hybrid data assimilation scheme where, we consider the 3DVAR scheme for the model as the constraint on the variational form, rather than constraining the variational form with the original model. We observe that this method reduces the computational cost of the minimization of the 4DVAR variational form, however, it introduces a bias in the estimate of the initial condition. We then explore how the results can be extended to weak constraint 4DVAR.

Chapter 1

Introduction

1.1 Overview

A large number of problems in applied mathematics are concerned with developing models with some predictive capability and with fine tuning of those models to obtain qualitative and quantitative insights into the physical dynamical systems. Most common application areas of such problems are oceanography, hydrology and numerical weather prediction (NWP) [13, 66], further applications can be found in [1, 33].

If the perfect model and corresponding initial conditions are known with certainty, the entire trajectory of the system under investigation can be computed. However, for most of the applications the models do not capture the physical phenomena completely due to the limitation on the theoretical understanding of the system or the presence of multiple random factors/parameters which can not be accounted for. Similarly, the initial conditions for the system may either be not known or available only as probabilistic estimates. These departures from the underlying system can cause large differences in estimation of actual state of the system and hamper the ability to predict the future states of the system.

In many of the applications the theoretical understanding of the system is complemented by the observations available for the system. The observations may have their own imperfections caused by imprecise measuring devices or inaccurate methodologies. Nonetheless, in many applications, the data can be incorporated with the model to get better estimates of the state of the system. The problem of state estimation from model forecasts and observed data can be formulated as a data assimilation problem [6, 10, 36].

The aim of data assimilation schemes is to obtain the estimate of the system as accurately as possible by combining the model and the observations [82, 81]. Since both of these sources are often erroneous, the estimate obtained *via* data assimilation process also contains errors and quantifying the uncertainty associated with the estimate is of crucial importance for a number of applications. At the same time many applications only require the estimate of the most probable state. To address these requirements two major approaches to data assimilation have evolved, the

first approach involves describing the system state as a probability distribution, whereas the other approach focuses solely on determining the optimal state of the system.

1.2 Models and observations

Before going into the description of the various data assimilation schemes, we establish the notation for the model and the observations.

1.2.1 Dynamical System

The model along with the prior estimates for the state of the system represent the theoretical knowledge available for the system under investigation. Let the vector $v_k \in \mathbb{R}^p$ be the state of the dynamical system to be estimated, and we model the evolution of the system by the following equation

$$u_k = \Psi_k(u_{k-1}),\ k \in \mathbb{Z}^+ \tag{1.2.1}$$

where the forward operator $\Psi_k(\cdot) : \mathbb{R}^p \to \mathbb{R}^p$ represents the computational model used to simulate the underlying system and map the estimate u_k forward in time.

If the forward operator Ψ_k captures the system dynamics perfectly, the estimate u_k can be evolved under the same equations. The aim in such situation is to estimate the initial condition v_0 accurately. However, if the underlying system is chaotic, the task of estimating the state becomes challenging as the small errors in estimate of the initial condition can introduce large errors over time.

When the underlying system is not known entirely the discrepancy between the system and the estimate can be modelled as stochastic inputs. Typically these errors arise from incorrect model formalisation, incorrect parameter values, numerical inaccuracies and rounding off errors. In the case when, model errors are present the forward model can be written as

$$u_k = \Psi_k(u_{k-1}) + \xi_k,\ k \in \mathbb{Z}^+ \tag{1.2.2}$$

where the term ξ_k accounts for the missing factors in the model. There are several formulations available to model the error term in forward equation as mentioned in section 1.4.2.

1.2.2 Observations

The observations collected on the system comprise of another source of information for the estimates to draw upon. In practice, it is often not possible to observe the system directly hence the following relation between the underlying system state v_k and the data $y \in \mathbb{R}^m$ is assumed

$$y_k = \mathcal{H}_k(v_k) + \nu_k\ k \in \mathbb{Z}^+, \tag{1.2.3}$$

where the term ν_k represents the random errors present in the observation process. The observation operator $\mathcal{H}_k(\cdot) : \mathbb{R}^p \to \mathbb{R}^m$ maps the system state to observation space. The observation noise process ν_k is assumed to be zero mean white noise process which does not depend upon the system state v_k.

In the next sections we give a brief introduction to the methodologies for sequential and variational data assimilation.

1.3 Sequential Data Assimilation

The uncertainty present both in the models available and the observed data opens the problem of estimating the state, to a probabilistic formulation. Under the probabilistic approach, the aim is to get the best estimate for the underlying system state as well as the uncertainty present in the estimation, in other words, one is looking to get the probability density function $P(u_k|Y_k)$ given the set of observations $Y_k = \{y_1, \ldots, y_k\}$. The distribution $P(u_k|Y_k)$ is called the filtering distribution.

Under the framework described in section 1.2, the Bayesian formulation provides a two-step iterative process for calculating the filtering distribution $P(u_k|Y_k)$. The first step is to propagate the state forward with the model and calculate the forecast distribution $P(u_k|Y_{k-1})$ as

$$P(u_k|Y_{k-1}) = \int P(u_k|u_{k-1})P(u_{k-1}|Y_{k-1})\mathrm{d}u_{k-1}$$

where the function $P(u_k|u_{k-1})$ depends on the model system. The second step is to use Bayes' theorem to update the forecast distribution to the filtering distribution as

$$P(u_k|Y_k) \propto P(y_k|u_k)P(u_k|Y_{k-1})$$

where the conditional distribution $P(y_k|u_k)$ is determined by the error statistics.

In the cases when the forward model and the observation operator are both linear and the model and the observation errors follow Gaussian distribution, the posterior distribution can be calculated explicitly. However, in most problems of interest the forward model and/or the observation operators are nonlinear which in general makes explicit calculations of the posterior distribution intractable. This intractability of posterior distribution has lead to the development of multiple approximation schemes some of which we describe in the following sections. Note that although Bayesian framework does allow for the error statistics to be non-Gaussian, most of the data assimilation schemes assume the Gaussian nature of the errors present in the model and the observations.

1.3.1 Kalman Filter

The Kalman filter is a sequential data assimilation scheme [35, 32] which provides optimal least square estimates for data assimilation problems. The statistical assumptions for the optimality

of the estimates are as following. Consider the case when model dynamics A_k and observation operator H_k are linear, given as

$$u_k = A_k u_{k-1} + \xi_k \tag{1.3.1}$$
$$y_k = H_k v_k + \nu_k, \tag{1.3.2}$$

best linear unbiased estimate can be obtained using Kalman Filter [36]. We also assume that the model error $\xi_k \sim N(0, \Sigma_k)$ and the observation error $\nu_k \sim N(0, \Gamma_k)$ follow Gaussian distributions. Further statistical assumptions include that the model error and the observation error processes are white noise processes, hence uncorrelated in time

$$\mathbb{E}[\xi_i^\top \xi_j] = \mathbb{E}[\nu_i^\top \nu_j] = 0, \forall i \neq j, \tag{1.3.3}$$

and are uncorrelated with each other

$$\mathbb{E}[\xi_i^\top \nu_i] = \mathbb{E}[\nu_j^\top \xi_j] = 0, \forall i, j. \tag{1.3.4}$$

The uncertainty in initial condition u_0 is taken care with assumption $u_0 \sim \mathcal{N}(u_0^a, P_0^a)$ where u_0^a is the guess for initial condition. To get the optimal estimate under the mentioned statistical assumptions Kalman Filter follows two step procedure, as described earlier, prediction step where the distribution of state is propagated forward in time followed by the assimilation step which updates the distribution of state given the newly available observation. In forecast step given the distribution of the state of the system $N(u_{k-1}^a, B_{k-1}^a)$ at time $k-1$, the forecast estimate u_k^f and associated covariance B_k^f are defined as

$$u_k^f = A_k u_{k-1}^a \tag{1.3.5}$$
$$B_k^f = A_k B_{k-1}^a A_k^\top + \Sigma_k. \tag{1.3.6}$$

If no data is present iterating above equations gives an estimate of the system state. When a new observation is available then the system's distribution is updated to assimilate new data into the forecast in the following way

$$u_k^a = (I - G_k H_k) u_k^f + G_k y_k \tag{1.3.7}$$
$$B_k^a = (I - G_k H_k) B_k^f, \tag{1.3.8}$$

where $G_k \in \mathbb{R}^{p \times m}$ is the Kalman gain matrix and it adjusts the relative weight of u_k^f and y_k. The Kalman gain matrix has the similar form as in the equation (1.3.24) where the forecast covariance serves as the background covariance

$$G_k = B_k^f H_k^\top (H_k B_k^f H_k^\top + \Gamma_k)^{-1}. \tag{1.3.9}$$

In cases where model dynamics and observation operator are linear, variance minimization can be done explicitly using Kalman Filter. However, when the system dynamics is non-linear, propagation of state error covariance matrix does not follow equation (1.3.6) and Kalman Filter does not give the optimal solution.

1.3.2 Extended Kalman Filter

For the systems where system dynamics and/or observation operator is non linear the estimation problem becomes intricate since unlike the linear case, the distribution of the states can not be completely characterized by second moments. Although, Kalman filtering algorithm can be generalized [32, 20] by linearizing the operators in neighbourhood of the forecast state u_k^f by computing the Jacobians of the dynamical model (1.2.2) and the observation operator (1.2.3)

$$A_k = \left.\frac{\partial \Psi_k}{\partial x}\right|_{u_k^f}, \quad H_k = \left.\frac{\partial \mathcal{H}_k}{\partial x}\right|_{u_k^f}$$

where the forecast is calculated by evolving the analysis step estimate u_{k-1}^a from the previous step according to the nonlinear model

$$u_k^f = \Psi_k(u_{k-1}^a), \tag{1.3.10}$$
$$B_k^f = A_k B_{k-1}^a A_k^\top + \Sigma_k. \tag{1.3.11}$$

The analysis step estimate is obtained by following the same steps as for the Kalman Filter but using nonlinear observation operator. The analysis estimate can be written as

$$u_k^a = u_k^f + G_k(y_k - \mathcal{H}_k(u_k^f)), \tag{1.3.12}$$
$$B_k^a = (I - G_k H_k) B_k^f, \tag{1.3.13}$$
$$G_k = B_k^f H_k^\top (H_k B_k^f H_k^\top + \Gamma_k)^{-1}. \tag{1.3.14}$$

where G_k is the Kalman Gain matrix. This scheme is called Extended Kalman Filter(EKF). Unlike the Kalman Filter the Extended Kalman Filter provides a suboptimal approximation of the state of the system. In the case of weakly non-linear systems, one can get good approximations from EKF. Whereas for highly non linear system dynamics and observation operators EKF does not perform well and sometimes leads to unstable approximations.

Implementing Kalman Filter or Extended Kalman Filter gives rise to two computational difficulties. The first is that the forecast covariance matrices B^f are required at each time step. But for most of the applications (e.g. geophysical, oceanic data) the size of forecast covariance matrix is of very large order ($10^{14} - 10^{18}$). Thus making the calculation of B^f computationally infeasible for real time state estimation. The second difficulty in implementing Kalman Filter or Extended Kalman Filter is finding the inverse of the matrix $\left(H_k B_k^f H_k^T + \Gamma_k\right)$ for each time step k. As previously, for practical applications, these are fairly large matrices and additionally if the

inverse is ill-conditioned it can make the solution be very sensitive to small errors.

To avoid these computational difficulties most operational sequential assimilation schemes consider approximations of above scheme. These are called sub-optimal or *ad-hoc* filtering schemes. In the next section we briefly describe one of such scheme, the Ensemble Kalman Filter, before moving on to variational schemes.

1.3.3 Ensemble Kalman Filter

The Ensemble Kalman Filter, first introduced in [17, 18], takes an alternative approach to estimating the background covariance matrix involves using Monte Carlo simulation. In Ensemble Kalman Filter a collection of state variables $\{u_0^i\}_{i\in\{1,\dots,N\}}$ is generated by sampling from the prior distribution and evolved following the system dynamics for each particle as

$$u_k^{i,f} = \Psi_k(u_{k-1}^{i,a}), \quad i \in \{1,\dots,N\} \tag{1.3.15}$$

and the forecast covariance matrix is approximated by the sample covariance matrix

$$\bar{u}_k^f = \frac{1}{N-1}\sum_{i=1}^{N} u_k^{i,f} \tag{1.3.16}$$

$$B_k^f = \frac{1}{N-1}\sum_{i=1}^{N}(u_k^{i,f} - \bar{u}_k^f)(u_k^{i,f} - \bar{u}_k^f)^\top \tag{1.3.17}$$

To update the state variable ensemble perturbed observations $\{z_k^i\}_{i\in\{1,\dots,N\}}$ are used. Given the observation y_k for the perturbed observations are generated from the distribution $z_k^i \sim N(y_k, \Gamma_k)$ for $i \in \{1,\dots,N\}$ as in [30]. The analysis step then can be written as

$$u_k^{i,a} = u_k^{i,f} + G_k(z_k^i - \mathcal{H}_k(u_k^{i,f})), \tag{1.3.18}$$

$$B_k^a = (I - G_k H_k)B_k^f, \tag{1.3.19}$$

$$G_k = B_k^f H_k^\top (H_k B_k^f H_k^\top + \Gamma_k)^{-1}. \tag{1.3.20}$$

The underlying assumption here is that the distribution of the ensemble particles appropriately captures the filtering distribution. The computational efficiency in formulation of EnKF algorithm has lead to its successful adoption into operational use. In practice, few of the main challenges in implementing this algorithm are, the sensitive dependence of the estimate on the initial ensemble and underestimation of the background covariance [24]. To address these issues many variants [25, 31, 9] have been proposed.

1.3.4 3DVAR

The 3DVAR algorithm [52, 12, 67] takes the approach of estimating the state of the system given the observation y_k and the forecast state/background state u_k^b at step k, by minimizing the following

cost function

$$J_k(u_k) = (u_k - u_k^b)^T {B_k}^{-1}(u_k - u_k^b) + (y_k - \mathcal{H}_k(u_k))^T {\Gamma_k}^{-1}(y_k - \mathcal{H}_k(u_k)) \tag{1.3.21}$$

where the covariance matrix B_k describes the covariances of the errors present in the forecast u_k^f whereas Γ_k captures the covariances present in the observation errors. The cost function $J_k(\cdot) : \mathbb{R}^p \to \mathbb{R}$ is designed to account for the effects of both the background state u_k^b and the observation y_k, weighted inversely by the uncertainty present as denoted by their covariance matrices respectively. If the background errors are smaller compared to the observation errors, the estimate obtained by minimizing the cost function is closer to the background and forecasts. Conversely, when the observation errors are small the assimilated state follows the observations [47].

In general, the assimilated state is obtained by minimizing the cost function given by the equation (1.3.21) as the observations become available, however, when the observation operator $\mathcal{H}_k$ is linear (represented as $H_k \in \mathbb{R}^{m\times p}$) the solution of the minimization problem

$$u_k^a = \underset{u_k}{\operatorname{argmin}}\ J_k(u_k) \tag{1.3.22}$$

can be obtained by the update step

$$u_k^a = u_k^b + G_k(y_k - H_k u_k^b) \tag{1.3.23}$$

where $G_k \in \mathbb{R}^{p\times m}$ is Kalman gain matrix, defined as

$$G_k = B_k H_k^\top (H_k B_k H_k^\top + \Gamma_k)^{-1}. \tag{1.3.24}$$

In the case when the observation operator is nonlinear a similar approach can be taken by approximating the observation operator by linearizing the observation operator as

$$H_k = \left.\frac{\partial \mathcal{H}}{\partial u}\right|_{u_k^b} \tag{1.3.25}$$

and substituting it in the update step as

$$u_k^a = u_k^b + G_k(y_k - \mathcal{H}_k u_k^b) \tag{1.3.26}$$

$$G_k = B_k H_k^\top (H_k B_k H_k^\top + \Gamma_k)^{-1}. \tag{1.3.27}$$

The key assumption in the application of 3DVAR algorithms is that over the period of assimilation the background error does not change significantly and the background error matrices B_k can be considered constant $B \in \mathbb{R}^{p\times p}$ for all assimilation steps. The update step is performed according to the equation (1.3.23) as described earlier. On one hand, the assumption of constant background covariance provides simplicity and efficiency to the 3DVAR assimilation scheme; how-

ever, it also is the biggest weakness of the assimilation scheme for the systems where forecast errors evolve rapidly. Another challenge in setting up 3DVAR assimilation scheme is the selection of appropriate background covariance matrix. More detailed discussions on these topics can be found in [62, 21].

1.4 4DVAR

The 3DVAR algorithm provides the estimate of the system state sequentially in time and the update step concerns only the most recent observation. Furthermore, in the optimization step only the predicted state is considered and the underlying model is not involved directly. The 4DVAR algorithm,first proposed by [45], extends the 3DVAR scheme temporally by accumulating the observations over a time window and minimizing over the model trajectory. 4DVAR methods can be classified in two categories, Strong Constraint and Weak Constraint, we describe both these formulations in the following sections.

1.4.1 Strong Constraint

Let us consider the system under investigation can be modeled by the following equation

$$u_k = \Psi_{0:k}(u_0) \tag{1.4.1}$$

where the variable u_k describes the system state and the solution operator $\Psi_{0:k}(\cdot) : \mathbb{R}^p \to \mathbb{R}^p$ maps the initial condition to the state at $k-$th step. Under the assumption that the model describes the underlying system perfectly i.e. if the true initial condition is known the complete trajectory of the system can be computed. However, in general instead of the true underlying initial condition one has the access to the estimate of the initial condition m_0 with the background covariance C_0. Let $Y_K = \{y_1, \ldots, y_K\}$ be the set of observations made over the assimilation window defined as

$$y_k = \mathcal{H}_k(v_k) + \nu_k,\ k \in \{1, \ldots, K\} \tag{1.4.2}$$

where the observation errors $\nu_k \sim N(0, \Gamma_k)$ are Gaussian. Similar to the cost function used for the 3DVAR algorithm the cost function in the 4DVAR algorithm also seeks to minimize the distance from the background estimate m_0 and the observations Y_K made over the assimilation window weighted by the inverse of the uncertainty present respectively. The objective function for the 4DVAR scheme given the set of observations over an assimilation time window has the following form

$$J(u_0) = (u_0 - m_0)^\top \, {C_0}^{-1} \, (u_0 - m_0) + \sum_{k=1}^{K} (y_k - \mathcal{H}_k(u_k))^\top {\Gamma_k}^{-1} (y_k - \mathcal{H}_k(u_k)). \tag{1.4.3}$$

The minimization of the objective function $J(\cdot) : \mathbb{R}^p \to \mathbb{R}$ under the constraint of the system dynamics provides the estimate for the initial condition of the system.

Although theoretically attractive the 4DVAR algorithm is computationally challenging for large nonlinear systems. The forward integration of the model at each minimization step makes it computationally intensive. Another problem is minimization process leading to a local minimum instead of global minimum.

1.4.2 Weak Constraint

One of the key assumptions in the formulation of strong constraint 4DVAR is that the model captures the underlying system dynamics perfectly. However, this assumption is rarely true in practice. One of the methods proposed to address model errors is weak constraint 4DVAR, first proposed in [69], where the model equations are modified to include model errors.

To take in the account of the imperfections of the model we modify the forward equation as

$$u_k = \Psi_k(u_{k-1}) + \eta_k \tag{1.4.4}$$

where η_k represents the model error present. The model errors are assumed to follow Gaussian distribution with zero mean and covariance matrix Σ_k as suggested in [26]. Although in this formulation of weak 4DVAR scheme we assume that the model errors are uncorrelated with the system state. However, in contrast to this assumption the area of systemic model errors has garnered a large amount of research interest [14]. Various formulations of accounting for model error have been explored in [44, 2]. Under the assumption of Gaussian model error the cost function (1.4.3) can be extended to include model errors as control variables which in turn reinforces the model as a weak constraint [83, 75]. The cost function for weak constraint 4DVAR can be given as

$$\begin{aligned} J\left(u_0, \{\eta_k\}_{k=1}^K\right) &= (u_0 - m_0)^\top C_0^{-1}(u_0 - m_0) + \sum_{k=1}^{K} (y_k - \mathcal{H}_k(u_k))^\top {\Gamma_k}^{-1}(y_k - \mathcal{H}_k(u_k)) \\ &+ \sum_{k=1}^{K} \eta_k^\top {\Sigma_k}^{-1}\eta_k. \end{aligned} \tag{1.4.5}$$

The final term penalizes the departure of the estimate from the model forecast. The cost function can alternatively be formulated as

$$\begin{aligned} J\left(\{u_k\}_{k=0}^K\right) &= (u_0 - m_0)^\top C_0^{-1}(u_0 - m_0) + \sum_{k=1}^{K} (y_k - \mathcal{H}_k(u_k))^\top {\Gamma_k}^{-1}(y_k - \mathcal{H}_k(u_k)) \\ &+ \sum_{k=1}^{K} (u_k - \Psi_k(u_{k-1}))^\top {\Sigma_k}^{-1}(u_k - \Psi_k(u_{k-1})). \end{aligned} \tag{1.4.6}$$

In the later formulation the model dynamics is assimilated in to the cost function explicitly. The

weak constrain formulation increases the dimension of minimization problem by the number of model integration steps leading to higher computational cost compared to the strong constraint formulation. Another challenge in setting up weak constraint 4DVAR scheme is approximation of model error statistics, in addition to the approximation of background covariance matrix as in both 3DVAR and strong constraint 4DVAR. To address these challenges multiple methods have been suggested, two prominent approaches are, restricting the background and model errors to the unstable and neutral subspace of the system [64, 72, 77] or using flow dependent error covariance matrices either computed by the linearized model or approximated *via* particle ensemble [51, 76, 49, 25]

1.5 Structure of the thesis

In Chapter 2 we study the application of the 3DVAR filtering scheme to the Lorenz'63 system when the observations are partial and contain random errors. We present theoretical results on the accuracy of the estimates for the case when assimilation happens in discrete time steps as well as in the case when the assimilation scheme has continuous formulation as introduced in [8]. We further corroborate the analytical results with the numerical experiments for both the discrete and continuous assimilation schemes.

In the first part of Chapter 3 we extend the application of the 3DVAR filtering scheme to the Lorenz'96 system. We derive the accuracy results for discrete and continuous assimilation formulation under suitable assumptions. The accuracy results for the continuous case are derived similarly to the results in Chapter 2, however, the discrete case results are more complex and requires additional assumptions on the structure of observational noise. In the second part of Chapter 3 we propose an adaptive observation scheme based on the linearized dynamics, which by focussing on the modes of maximal growth reduces the required number of observation for accurate estimation. We perform numerical experiments for Lorenz'96 system with the proposed observation operator. The numerical experiments are performed for both the 3DVAR and the Extended Kalman Filter schemes with partial and noisy observations.We also draw connections to the work on assimilation in the unstable space approach presented in [78]. In Chapter 4 we move from sequential data assimilation schemes to variational data assimilation scheme. We again consider a data assimilation system where observations are noisy, however the aim of the analysis in this chapter, is to improve the efficiency of minimization at the cost of introducing a bias in the estimate. In this chapter we introduce a hybrid scheme based on the strong constraint 4DVAR formulation where the trajectory of the model is constrained by the 3DVAR estimates instead of the system dynamics. We present analytical and numerical results on the presence of bias under the 3DVAR constrained 4DVAR scheme for unstable systems. We further extend the numerical study to the Lorenz'63 and the Lorenz'96 models which are nonlinear and chaotic.

Chapter 5 expands the ideas from previous chapter to weak 4DVAR scheme. The 3DVAR constrained 4DVAR scheme discussed in the the weak formulation of the 4DVAR cost functional.

The 3DVAR estimate is used as a weak dynamical constraint. We establish analytical bounds on the error introduced in the estimate in case of linear model. We also demonstrate the numerical results pertaining to the Lorenz'63 model.

Chapter 2

Partial observations on Lorenz'63 system

2.1 Introduction

Data assimilation concerns estimation of the state of a dynamical system by combining observed data with the underlying mathematical model. It finds widespread application in the geophysical sciences, including meteorology [36], oceanography [6] and oil reservoir simulation [59]. Both filtering methods, which update the state sequentially, and variational methods, which can use an entire time window of data, are used [3]. However, the dimensions of the systems arising in the applications of interest are enormous – of $\mathcal{O}(10^9)$ in global weather forecasting, for example. This makes rigorous Bayesian approaches such as the sequential particle filter [16], for the filtering problem, or MCMC methods for the variational problem [71], prohibitively expensive in on-line scenarios.

For this reason various *ad hoc* methodologies are typically used. In the context of filtering these usually rely on making some form of Gaussian ansatz [80]. The 3DVAR method [50, 62] is the simplest Gaussian filter, relying on fixed (with respect to the data time-index increment) forecast and analysis model covariances, related through a Kalman update. A more sophisticated idea is to update the forecast covariance via the linearized dynamics, again computing the analysis covariance via a Kalman update, leading to the extended Kalman filter [32]. In high dimensions computing the full linearized dynamics is not practical. For this reason the ensemble Kalman filter [19, 20] is widely used, in which the forecast covariance is estimated from an ensemble of particles, and each particle is updated in Kalman fashion. An active current area of research in filtering concerns the development of methods which retain the computational expediency of approximate Gaussian filters, but which incorporate physically motivated structure into the forecast and analysis steps [57, 56], and are non-Gaussian.

Despite the widespread use of these many variants on approximate Gaussian filters, systematic mathematical analysis remains in its infancy. Because the 3DVAR method is prototypical of other more sophisticated *ad hoc* filters it is natural to develop a thorough understanding of the

mathematical properties of this filter. Two recent papers address these issues in the context of the Navier-Stokes equation, for data streams which are discrete in time [11] and continuous in time [8]. These papers study the situation where the observations are partial (only low frequency spatial information is observed) and subject to small noise. Conditions are established under which the filter can recover from an order one initial error and, after enough time has elapsed, estimate the entire system state to within an accuracy level determined by the observational noise scale; this is termed *filter accuracy*. Key to understanding, and proving, these results on the 3DVAR filter for the Navier-Stokes equation are a pair of papers by Titi and co-workers which study the synchronization of the Navier-Stokes equation with a true signal which is fed into only the low frequency spatial modes of the system, without noise [60, 29]; the higher modes then synchronize because of the underlying dynamics. The idea that a finite amount of information effectively governs the large-time behaviour of the Navier-Stokes equation goes back to early studies of the equation as a dynamical system [23] and is known as the *determining node* or *mode* property in the modern literature [68]. The papers [11, 8] demonstrate that the technique of *variance inflation*, widely employed by practitioners in high dimensional filtering, can be understood as a method to add greater weight to the data, thereby allowing the synchronization effect to take hold.

The Lorenz '63 model [53, 70] provides a useful metaphor for various aspects of the Navier-Stokes equation, being dissipative with a quadratic energy-conserving nonlinearity [22]. In particular, the Lorenz model exhibits a form of synchronization analogous to that mentioned above for the Navier-Stokes equation [29]. This strongly suggests that results proved for 3DVAR applied to the Navier-Stokes equation will have analogies for the Lorenz equations. The purpose of this paper is to substantiate this assertion.

The presentation is organized as follows. In section 2.2 we describe the Bayesian formulation of the inverse problem of sequential data assimilation; we also present a brief introduction to the relevant properties of the Lorenz '63 model and describe the 3DVAR filtering schemes for both discrete and continuous time data streams. In section 2.3 we derive Theorem 2.3.2 concerning the 3DVAR algorithm applied to the Lorenz model with discrete time data. This is analogous to Theorem 3.3 in [11] for the Navier-Stokes equation. However, in contrast to that paper, we study Gaussian (and hence unbounded) observational noise and, as a consequence, our results are proved in mean square rather than almost surely. In section 2.4 we extend the accuracy result to the continuous time data stream setting: Theorem 2.4.1; the result is analogous to Theorem 4.3 in [8] which concerns the Navier-Stokes equation. Section 2.5 contains numerical results which illustrate the theory. We make concluding remarks in section 2.6.

2.2 Set-Up

In subsection 2.2.1 we formulate the probabilistic inverse problem which arises from attempting to estimate the state of a dynamical system subject to uncertain initial condition, and given partial, noisy observations. Subsection 2.2.2 introduces the Lorenz '63 model which we employ throughout

this paper. In subsections 2.2.3 and 2.2.4 we describe the discrete and continuous 3DVAR filters whose properties we study in subsequent sections.

2.2.1 Inverse Problem

Consider a model whose dynamics is governed by the equation

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \mathcal{F}(u), \tag{2.2.1}$$

with initial condition $u(0) = u_0 \in \mathbb{R}^p$. We assume the the initial condition is uncertain and only its statistical distribution is known, namely the Gaussian $u_0 \sim N(m_0, C_0)$. Assuming that the equation has a solution for any $u_0 \in \mathbb{R}^p$ and all positive times, we let $\Psi(\cdot,\cdot) : \mathbb{R}^p \times \mathbb{R}^+ \to \mathbb{R}^p$ be the solution operator for equation (2.2.1). Now suppose that we observe the system at equally spaced times $t_k = kh$ for all $k \in \mathbb{Z}^+$. For simplicity we write $\Psi(\cdot) := \Psi(\cdot; h)$. Defining $u_k = u(t_k) = \Psi(u_0; kh)$ we have

$$u_{k+1} = \Psi(u_k), \quad k \in \mathbb{Z}^+. \tag{2.2.2}$$

We assume that the data $\{y_k\}_{k\in\mathbb{Z}^+}$ is found from noisily observing a linear operator H applied to the system state, at each time t_k, so that

$$y_{k+1} = Hu_{k+1} + \nu_{k+1}, \quad k \in \mathbb{N}. \tag{2.2.3}$$

Here $\{\nu_k\}_{k\in\mathbb{N}}$ is an i.i.d. sequence of random variables, independent of u_0, with $\nu_1 \sim N(0, \Gamma)$ and H denotes a linear operator from $\mathbb{R}^p$ to $\mathbb{R}^m$, with $m \leq p$. If the rank of H is less than p the system is said to be *partially observed*. The partially observed situation is the most commonly arising in applications and we concentrate on it here. The over-determined case $m > p$ corresponds to making more than one observation in certain directions; one approach that can be used in this situation is to average multiple observations to reduce the effective observational error variance by the square root of the number of observations in that direction, and thereby reduce to the case where the rank is less than or equal to p.

We denote the accumulated data up to time k by $Y_k := \{y_j\}_{j=1}^k$. The pair (u_k, Y_k) is a jointly varying random variable in $\mathbb{R}^p \times \mathbb{R}^{km}$. The goal of filtering is to determine the distribution of the conditioned random variable $u_k|Y_k$, and to update it sequentially as k is incremented. This corresponds to a sequence of inverse problems for the system state, given observed data, and it has been regularized by means of the Bayesian formulation.

2.2.2 Forward Model: Lorenz '63

When analyzing the 3DVAR approach to the filtering problem we will focus our attention on a particular model problem, namely the classical Lorenz '63 system [53]. In this section we introduce the model and summarize the properties relevant to this paper. The Lorenz equations are a system

of three coupled non-linear ordinary differential equations whose solution $u \in \mathbb{R}^3$, where $u = (u_x, u_y, u_z)$, satisfies

$$\dot{u}_x = \alpha(u_y - u_x), \tag{2.2.4a}$$
$$\dot{u}_y = -\alpha u_x - u_y - u_x u_z, \tag{2.2.4b}$$
$$\dot{u}_z = u_x u_y - b u_z - b(r + \alpha). \tag{2.2.4c}$$

Note that we have employed a coordinate system where origin is shifted to the point $\big(0, 0, -(r+\alpha)\big)$ as discussed in [74]. Throughout this paper we will use the classical parameter values $(\alpha, b, r) = (10, \frac{8}{3}, 28)$ in all of our numerical experiments. At these values, the system is chaotic [79] and has one positive and one negative Lyapunov exponent and the third is zero, reflecting time translation-invariance. Our theoretical results, however, simply require that $\alpha, b > 1$ and $r > 0$ and we make this assumption, without further comment, throughout the remainder of the paper.

In the following it is helpful to write the Lorenz equation in the following form as given in [22],[29]:

$$\frac{du}{dt} + Au + B(u, u) = f, \quad u(0) = u_0, \tag{2.2.5}$$

where

$$A = \begin{pmatrix} \alpha & -\alpha & 0 \\ \alpha & 1 & 0 \\ 0 & 0 & b \end{pmatrix}, \quad f = \begin{pmatrix} 0 \\ 0 \\ -b(r+\alpha) \end{pmatrix}$$

$$B(u, \tilde{u}) = \begin{pmatrix} 0 \\ (u_x \tilde{u_z} + u_z \tilde{u_x})/2 \\ -(u_x \tilde{u_y} + u_y \tilde{u_x})/2 \end{pmatrix}.$$

We use the notation $\langle \cdot, \cdot \rangle$ and $|\cdot|$ for the standard Euclidean inner-product and norm. When describing our observations it will also be useful to employ the projection matrices P and Q defined by

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad Q = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{2.2.6}$$

Remark 2.2.1. *Note that in this work we have chosen the x-component as the observed variable. We could have also chosen the y-component as the observed variable to obtain the convergence although the analytical results in that case resemble the results from next chapter. Choosing the z-component is not sufficient for convergence, more details can be found in [63].*

We will use the following properties of A and B:

Properties 2.2.2 ([29])**.** *For all $u, \tilde{u} \in \mathbb{R}^3$*

1. $\langle Au, u \rangle = \alpha u_x^2 + u_y^2 + b u_z^2 > |u|^2$ *provided that* $\alpha, b > 1$.

2. $\langle B(u,u),u\rangle = 0$.

3. $B(u,\tilde{u}) = B(\tilde{u},u)$.

4. $|B(u,\tilde{u})| \leq 2^{-1}|u||\tilde{u}|$.

5. $|\langle B(u,\tilde{u}),\tilde{u}\rangle| \leq 2^{-1}|u||\tilde{u}||P\tilde{u}|$.

We will also use the following:

Proposition 2.2.3. *([29], Theorem 2.2) Equation (2.2.5) has a global attractor $\mathcal{A}$. Let u be a trajectory with $u_0 \in \mathcal{A}$. Then $|u(t)|^2 \leq K$ for all $t \in \mathbb{R}$ where*

$$K = \frac{b^2(r+\alpha)^2}{4(b-1)}. \tag{2.2.7}$$

Figure 2.2.1 illustrates the properties of the equation. Sub-figure 2.2.1a shows the global attractor $\mathcal{A}$. Sub-figures 2.2.1b, 2.2.1c and 2.2.1d show the components u_x, u_y and u_z, respectively, plotted against time.

2.2.3 3DVAR: Discrete Time Data

In this section we describe the 3DVAR filtering scheme for the model (2.2.1) in the case where the system is observed discretely at equally spaced time points. The system state at time $t_k = kh$ is denoted by $u_k = u(t_k)$ and the data upto that time is $Y_k = \{y_j\}_{j=1}^k$. Recall that our aim is to find the probability distribution of $u_k|Y_k$. Approximate Gaussian filters, of which 3DVAR is a prototype, impose the following approximation:

$$\mathbb{P}(u_k|Y_k) = N(m_k, C_k). \tag{2.2.8}$$

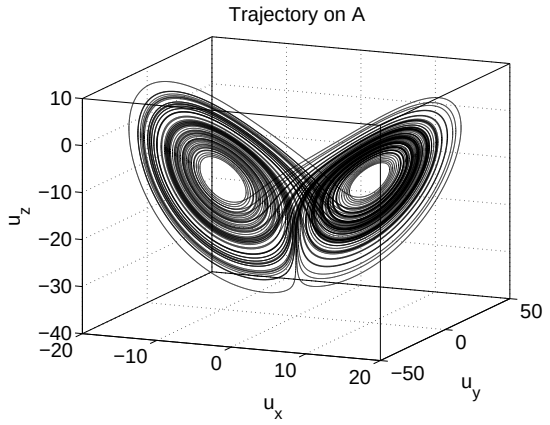
Given this assumption the filtering scheme can be written as an update rule

$$(m_k, C_k) \mapsto (m_{k+1}, C_{k+1}). \tag{2.2.9}$$

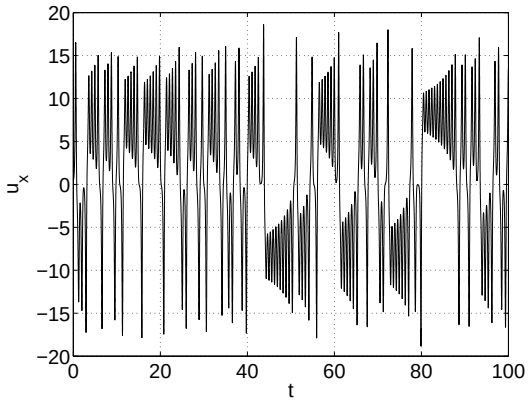
To determine this update we make a further Gaussian approximation, namely that u_{k+1} given Y_k follows a Gaussian distribution:

$$\mathbb{P}(u_{k+1}|Y_k) = N(\hat{m}_{k+1}, \hat{C}_{k+1}). \tag{2.2.10}$$

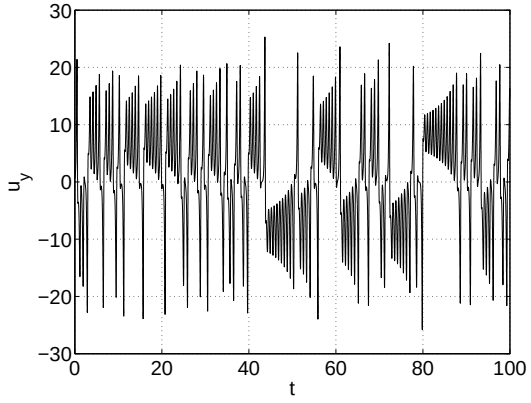
Now we can break the update rule into two steps of *prediction* $(m_k, C_k) \mapsto (\hat{m}_{k+1}, \hat{C}_{k+1})$ and *analysis* $(\hat{m}_{k+1}, \hat{C}_{k+1}) \to (m_{k+1}, C_{k+1})$. For the prediction step we assume that $\hat{m}_{k+1} = \Psi(m_k)$ whilst the choice of the covariance matrix $\hat{C}_{k+1}$ depends upon the choice of particular approximate Gaussian filter under consideration. For the analysis step, (2.2.10) together with the fact that



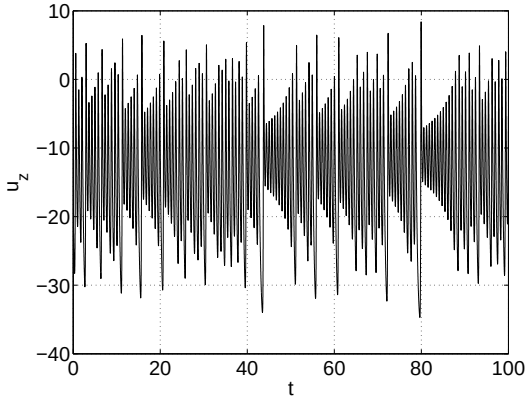
(a) Sample trajectory for a solution lying on global attractor.



(b) u_x component



(c) u_y component



(d) u_z component

Figure 2.2.1: Lorenz attractor and individual components.

$y_{k+1}|u_{k+1} \sim N(Hu_{k+1}, \Gamma)$ and application of Bayes' rule, implies that

$$u_{k+1}|Y_{k+1} \sim N(m_{k+1}, C_{k+1}) \tag{2.2.11}$$

where [28]

$$\begin{aligned} C_{k+1} &= \hat{C}_{k+1} - \hat{C}_{k+1}H^*(\Gamma + H\hat{C}_{k+1}H^*)^{-1}H\hat{C}_{k+1} && \text{(2.2.12a)} \\ m_{k+1} &= \Psi(m_k) + \hat{C}_{k+1}H^*(\Gamma + H\hat{C}_{k+1}H^*)^{-1}\big(y_{k+1} - H\Psi(m_k)\big). && \text{(2.2.12b)} \end{aligned}$$

As mentioned the choice of update rule $C_k \to \hat{C}_{k+1}$ defines the particular approximate Gaussian filtering scheme. For the 3DVAR scheme we impose $\hat{C}_{k+1} = C \quad \forall k \in \mathbb{N}$ where C is a positive definite $p \times p$ matrix. From equation (2.2.12b) we then get

$$\begin{aligned} m_{k+1} &= \Psi(m_k) + CH^*(\Gamma + HCH^*)^{-1}\big(y_{k+1} - H\Psi(m_k)\big) \\ &= (I - GH)\Psi(m_k) + Gy_{k+1} && \text{(2.2.13)} \end{aligned}$$

where

$$G := CH^*(\Gamma + HCH^*)^{-1} \tag{2.2.14}$$

is called Kalman gain matrix. The iteration (2.2.13) is analyzed in section 2.3.

Another way of defining the 3DVAR filter is by means of the following variational definition:

$$m_{k+1} = \underset{m}{\operatorname{argmin}} \left(\frac{1}{2} \| C^{-\frac{1}{2}} \big(m - \Psi(m_k) \big) \|^2 + \frac{1}{2} \| \Gamma^{-\frac{1}{2}} (y_{k+1} - Hm) \|^2 \right). \tag{2.2.15}$$

This coincides with the previous definition because the mean of a Gaussian can be characterized as the minimizer of the negative of the logarithm of the probability density function and because the analysis step corresponds to a Bayesian Gaussian update, given the assumptions underlying the filter; indeed the fact that the negative logarithm is the sum of two squares follows from Bayes' theorem. From the variational formulation, it is clear that the 3DVAR filter is a compromise between fitting the model and the data. The model uncertainty is characterized by a fixed covariance C, and the data uncertainty by a fixed covariance Γ; the ratio of the size of these two covariances will play an important role in what follows.

2.2.4 3DVAR: Continuous Time Data

In this section we describe the limit of high frequency observations $h \to 0$ which, with appropriate scaling of the noise covariance with respect to the observation time h, leads to a stochastic differential equation (SDE) limit for the 3DVAR filter. We refer to this SDE as the continuous time 3DVAR filter. We give a brief derivation, referring to [8] for further details and to [7] for a related analysis of continuous time limits in the context of the ensemble Kalman filter.

We assume the following scaling for the observation error covariance matrix: $\Gamma = \frac{1}{h}\Gamma_0$. Thus, although the data arrives more and more frequently, as we consider the limit $h \to 0$, it is also becoming more uncertain; this trade-off leads to the SDE limit. Define the sequence of variables $\{z_k\}_{k\in\mathbb{N}}$ by the relation $z_{k+1} = z_k + hy_{k+1}$ and $z_0 = 0$. Then

$$z_{k+1} = z_k + hHu_{k+1} + \sqrt{h\Gamma_0}\gamma_k, \quad z_0 = 0. \tag{2.2.16}$$

Here $\gamma_k \sim N(0, I)$. By rearranging and taking limit as $h \to 0$ we get

$$\frac{\mathrm{d}z}{\mathrm{d}t} = Hu + \sqrt{\Gamma_0}\frac{\mathrm{d}w}{\mathrm{d}t}, \tag{2.2.17}$$

where w is an $\mathbb{R}^m$ valued standard Brownian motion. We think of $Z(t) := \{z(s)\}_{s\in[0,t]}$ as being the data. For each fixed t we have the jointly varying random variable $(u(t), Z(t)) \in \mathbb{R}^p \times C([0,t];\mathbb{R}^m)$. We are interested in the filtering problem of determining the sequence of conditioned probability distributions implied by the random variable $u(t)|Z(t)$ in $\mathbb{R}^p$. The 3DVAR filter imposes Gaussian approximations of the form $N\big(m(t), C\big)$. We now derive the evolution equation for $m(t)$.

Recall the vector field $\mathcal{F}$ which drives equation (2.2.1). Using equation (2.2.16) in (2.2.13),

together with the fact that $\Psi(u) = u + h\mathcal{F}(u) + \mathcal{O}(h^2)$, gives

$$m_{n+1} = m_n + h\mathcal{F}(m_n) + \mathcal{O}(h^2) + hCH^*(\Gamma_0 + hHCH^*)^{-1}\left(\frac{z_{n+1} - z_n}{h} - Hm_n\right). \tag{2.2.18}$$

Rearranging and taking limit $h \to 0$ gives

$$\frac{\mathrm{d}m}{\mathrm{d}t} = \mathcal{F}(m) + CH^*\Gamma_0^{-1}\left(\frac{\mathrm{d}z}{\mathrm{d}t} - Hm\right). \tag{2.2.19}$$

Equation (2.2.19) defines the continuous time 3DVAR filtering scheme and is analyzed in section 2.4. The data should be viewed as the continuous time stream $Z(t) = \{z(s)\}_{s\in[0,t]}$ and equations (2.2.17) and (2.2.19) as stochastic differential equations driven by w and z respectively.

2.3 Analysis of Discrete Time 3DVAR

In this section we analyse the discrete time 3DVAR algorithm when applied to a partially observed Lorenz '63 model; in particular we assume only that the u_x component is observed. We start, in subsection 2.3.1, with some general discussion of error propagation properties of the filter. In subsection 2.3.2 we study mean square behaviour of the filter for Gaussian noise. Recall the projection matrices P and Q given by (2.2.6), we will use these in the following. We will also use $\{v_k\}$ to denote the exact solution sequence from the Lorenz equations which underlies the data; this is to be contrasted with $\{u_k\}$ which denotes the random variable which, when conditioned on the data, is approximated by the 3DVAR filter.

2.3.1 Preliminary Calculations

Throughout we assume that $H = (1, 0, 0)$, so that only u_x is observed, and we choose the model covariance $C = \eta^{-1}\epsilon^2 I$. We also assume that $\Gamma = \epsilon^2$. The Kalman gain matrix is then $G = \frac{1}{1+\eta}H^*$ and the 3DVAR filter (2.2.13) may be written

$$m_{k+1} = \left(\frac{\eta}{1+\eta}P + Q\right)\Psi(m_k) + \frac{1}{1+\eta}y_{k+1}H^*. \tag{2.3.1}$$

The scalar parameter is a design parameter whose choice we will discuss through the analysis of the iteration (2.3.1). Note that we are working with rather specific choices of model and observational noise covariances C and Γ; we will comment on generalizations in the concluding section 2.6.

We define v to be the true solution of the Lorenz equation (2.2.5) which underlies the data, and we define $v_k = v(kh)$, the solution at observation times. Note that, since $\Gamma = \epsilon^2$, it is consistent

to assume that the observation errors have the form

$$\nu_k = \begin{pmatrix} \epsilon\xi_k \\ 0 \\ 0 \end{pmatrix},$$

where ξ_k are i.i.d. random variables on $\mathbb{R}$. We will consider the case $\xi_1 \sim N(0,1)$ for simplicity of exposition. Note that we may write

$$\begin{aligned} y_{k+1}H^* &= Pv_{k+1} + \nu_{k+1} \\ &= P\Psi(v_k) + \nu_{k+1}. \end{aligned}$$

Thus

$$m_{k+1} = \left(\frac{\eta}{1+\eta}P + Q\right)\Psi(m_k) + \frac{1}{1+\eta}\Big(P\Psi(v_k) + \nu_{k+1}\Big). \tag{2.3.2}$$

Observe that

$$v_{k+1} = \Psi(v_k) = \left(\frac{\eta}{1+\eta}P + Q\right)\Psi(v_k) + \frac{1}{1+\eta}P\Psi(v_k). \tag{2.3.3}$$

We are interested in comparing m_k, the output of the filter, with v_k the true signal which underlies the data. We define the error process $\delta(t)$ as follows:

$$\delta(t) = \begin{cases} m_k - v(t) & \text{if } t = t_k \\ \Psi(m_k, t - t_k) - v(t) & \text{if } t \in (t_k, t_{k+1}) \end{cases}$$

Observe that δ is discontinuous at times t_j which are multiples of h, since $m_{k+1} \neq \Psi(m_k; h)$. In the following we write $\delta(t_j^-)$ for $\lim_{t\to t_j^-}\delta(t)$ and we define $\delta_j = \delta(t_j)$. Thus $\delta_j \neq \delta(t_j^-)$. Subtracting (2.3.3) from (2.3.2) we obtain

$$\delta(t_{k+1}) = \left(\frac{\eta}{1+\eta}P + Q\right)\delta(t_{k+1}^-) + \frac{1}{1+\eta}\nu_k. \tag{2.3.4}$$

Now consider the time interval (t_k, t_{k+1}). Since $\delta(t)$ is simply given by the difference of two solutions of the Lorenz equations in this interval, we have

$$\frac{d\delta}{dt} + A\delta + B(v,\delta) + B(\delta,v) + B(\delta,\delta) = 0, \quad t \in (t_k, t_{k+1}). \tag{2.3.5}$$

Taking the Euclidean inner product of equation (2.3.5) with δ gives

$$\frac{1}{2}\frac{d|\delta|^2}{dt} + \langle A\delta, \delta\rangle + \langle B(v,\delta),\delta\rangle + \langle B(\delta,v),\delta\rangle + \langle B(\delta,\delta),\delta\rangle = 0 \tag{2.3.6}$$

which, on simplifying and using Properties 2.2.2, gives

$$\frac{1}{2}\frac{d|\delta|^2}{dt} + \langle A\delta, \delta\rangle + 2\langle B(v,\delta),\delta\rangle = 0, \tag{2.3.7}$$

and hence

$$\frac{1}{2}\frac{d|\delta|^2}{dt} + |\delta|^2 + 2\langle B(v,\delta),\delta\rangle \leq 0. \tag{2.3.8}$$

In order to use (2.3.4) we wish to estimate the behaviour of $\delta(t_{k+1}^-)$ in terms of δ_k. The following is useful in this regard and may be proved by using (2.3.8) together with Properties 2.2.2(4). Note that K is defined by equation (2.2.7) and is necessarily greater than or equal to one, since $b, \alpha > 1$.

Proposition 2.3.1 ([29]). *Assume the true solution v lies on the global attractor $\mathcal{A}$ so that $\sup_{t\geq 0}|v(t)|^2 \leq K$ with*

$$K = \frac{b^2(r+\alpha)^2}{4(b-1)}.$$

Then for $\beta = 2\left(K^{1/2} - 1\right)$ it follows that $|\delta(t)|^2 \leq |\delta_k|^2 e^{\beta(t-t_k)}$ for $t \in [t_k, t_{k+1})$.

2.3.2 Accuracy Theorem

In this subsection we assume that $\xi_1 \sim N(0,1)$ and we study the behaviour of the filter in forward time when the size of the observational noise, $\mathcal{O}(\epsilon)$, is small. The following result shows that, provided variance inflation is employed (η small enough), the 3DVAR filter can recover from an $\mathcal{O}(1)$ initial error and enter an $\mathcal{O}(\epsilon)$ neighbourhood of the true signal. The results are proved in mean square. The reader will observe that the bound on the error behaves poorly as the observation time h goes to zero, a result of the over-weighting of observed data which is fluctuating wildly as $h \to 0$. This effect is removed in section 2.4 where the observational noise is scaled appropriately, in terms of $h \to 0$, to avoid this effect.

For this theorem we define a norm $\|\cdot\|$ by $\|u\|^2 = |u|^2 + |Pu|^2$, where $|\cdot|$ is the Euclidean norm.

Theorem 2.3.2. *Let v be a solution of the Lorenz equation (2.2.5) with $v(0) \in \mathcal{A}$, the global attractor. Assume that $\xi_1 \sim N(0,1)$ so that the observational noise is Gaussian. Then there exist $h_c > 0$, $\lambda > 0$ such that for all η sufficiently small and all $h \in (0, h_c)$*

$$\mathbb{E}||\delta_{k+1}||^2 \leq (1-\lambda h)\mathbb{E}||\delta_k||^2 + 2\epsilon^2. \tag{2.3.9}$$

Consequently

$$\limsup_{k\to\infty} \mathbb{E}||\delta_k||^2 \leq \frac{2\epsilon^2}{\lambda h}. \tag{2.3.10}$$

Proof. Recall that we have $\mathbb{E}\nu_{k+1} = 0$ and $\mathbb{E}|\nu_{k+1}|^2 = \epsilon^2$. On application of the projection P to

the error equation (2.3.4) for 3DVAR we obtain

$$\mathbb{E}|P\delta_{k+1}|^2 \leq \left(\frac{\eta}{1+\eta}\right)^2 \mathbb{E}|P\delta(t_{k+1}^-)|^2 + \left(\frac{1}{1+\eta}\right)^2 \epsilon^2. \tag{2.3.11}$$

Since $\mathbb{E}|Q\delta_{k+1}|^2 = \mathbb{E}|Q\delta(t_{k+1}^-)|^2 \leq \mathbb{E}|\delta(t_{k+1}^-)|^2$ we also obtain the bound

$$\mathbb{E}|\delta_{k+1}|^2 \leq \left(\frac{\eta}{1+\eta}\right)^2 \mathbb{E}|P\delta(t_{k+1}^-)|^2 + \mathbb{E}|\delta(t_{k+1}^-)|^2 + \left(\frac{1}{1+\eta}\right)^2 \epsilon^2. \tag{2.3.12}$$

Define M_1 and M_2 by

$$M_1(\tau) = \frac{K\alpha}{\beta+\alpha}\left(\frac{e^{\beta\tau}-e^{-\tau}}{\beta+1} - \frac{e^{-\alpha\tau}-e^{-\tau}}{1-\alpha}\right) + e^{-\tau} + 2\left(\frac{\eta}{1+\eta}\right)^2 \left(\frac{\alpha}{\beta+\alpha}\right)(e^{\beta\tau}-e^{-\alpha\tau}) \tag{2.3.13}$$

and

$$M_2(\tau) = \frac{K}{1-\alpha}\left(e^{-\alpha\tau}-e^{-\tau}\right) + 2\left(\frac{\eta}{1+\eta}\right)^2 e^{-\alpha\tau}. \tag{2.3.14}$$

Adding (2.3.11) to (2.3.12) and using Lemma 2.3.3 shows that

$$\mathbb{E}\|\delta_{k+1}\|^2 \leq M_1(h)\mathbb{E}|\delta_k|^2 + M_2(h)\mathbb{E}|P\delta_k|^2 + 2\left(\frac{1}{1+\eta}\right)^2 \epsilon^2, \tag{2.3.15}$$

so that

$$\mathbb{E}||\delta_{k+1}||^2 \leq M(h)\mathbb{E}||\delta_k||^2 + \frac{2\epsilon^2}{(1+\eta)^2}, \tag{2.3.16}$$

where

$$M(\tau) = \max\{M_1(\tau), M_2(\tau)\}. \tag{2.3.17}$$

Now we observe that

$$M_1(0) = 1,\ M_1'(0) = -1 + 2\alpha\left(\frac{\eta}{1+\eta}\right)^2 \text{ and } M_2(0) = 2\left(\frac{\eta}{1+\eta}\right)^2.$$

Thus there exists an $h_c > 0$ and a $\lambda > 0$ such that, for all η sufficiently small

$$M(\tau,\eta) \leq 1 - \lambda\tau, \quad \forall \tau \in (0, h_c].$$

Hence the theorem is proved. □

The following lemma is used in the preceding proof.

Lemma 2.3.3. *Under the conditions of Theorem 2.3.2 for $t \in [t_k, t_{k+1})$ we have*

$$|P\delta(t)|^2 \leq \frac{\alpha|\delta_k|^2}{\beta+\alpha}\left(e^{\beta(t-t_k)} - e^{-\alpha(t-t_k)}\right) + |P\delta_k|^2 e^{-\alpha(t-t_k)} \tag{2.3.18}$$

and

$$\begin{aligned}|\delta(t)|^2 \leq & \frac{K\alpha|\delta_k|^2}{\beta+\alpha}\left(\frac{e^{\beta(t-t_k)}-e^{-(t-t_k)}}{\beta+1}-\frac{e^{-\alpha(t-t_k)}-e^{-(t-t_k)}}{1-\alpha}\right)\\ &+\frac{K|P\delta_k|^2}{1-\alpha}\left(e^{-\alpha(t-t_k)}-e^{-(t-t_k)}\right)+|\delta_k|^2e^{-(t-t_k)}.\end{aligned} \tag{2.3.19}$$

Proof. Taking inner product of (2.3.5) with $P\delta$, instead of with δ as previously, we get

$$\frac{1}{2}\frac{d|P\delta|^2}{dt}+\langle A\delta, P\delta\rangle = 0. \tag{2.3.20}$$

Let $\delta = (\delta_x, \delta_y, \delta_z)^T$. Notice that $|P\delta|^2 = |\delta_x|^2$ and $\langle A\delta, P\delta\rangle = \alpha\delta_x^2 - \alpha\delta_x\delta_y$. Therefore equation (2.3.20) becomes

$$\begin{aligned}\frac{1}{2}\frac{d|P\delta|^2}{dt}+\alpha\delta_x^2 &= \alpha\delta_x\delta_y\\ &\leq \frac{\alpha}{2}\delta_x^2+\frac{\alpha}{2}\delta_y^2\\ &\leq \frac{\alpha}{2}\delta_x^2+\frac{\alpha}{2}|\delta|^2.\end{aligned}$$

By rearranging and applying Proposition 2.3.1 we get

$$\frac{d|P\delta|^2}{dt}+\alpha|P\delta|^2 \leq \alpha|\delta(t_k)|^2e^{\beta(t-t_k)}. \tag{2.3.21}$$

Multiplying by integrating factor $e^{\alpha(t-t_k)}$ and integrating from t_k to t gives equation (2.3.18).

Analysing the non-linear term in equation (2.3.8) with Property 2.2.2(5) gives

$$\begin{aligned}|2\langle B(v,\delta),\delta\rangle| &\leq |v||P\delta||\delta|\\ &\leq K^{\frac{1}{2}}|P\delta||\delta|\\ &\leq \frac{1}{2}K|P\delta|^2+\frac{1}{2}|\delta|^2.\end{aligned} \tag{2.3.22}$$

Substituting (2.3.18) and (2.3.22) in (2.3.8) gives

$$\frac{d|\delta|^2}{dt}+|\delta|^2 \leq \frac{K\alpha|\delta_k|^2}{\beta+\alpha}\left(e^{\beta(t-t_k)}-e^{-\alpha(t-t_k)}\right)+K|P\delta_k|^2e^{-\alpha(t-t_k)}. \tag{2.3.23}$$

Multiplying by the integrating factor $e^{(t-t_k)}$ and integrating from t_k to t gives

$$|\delta(t)|^2e^{(t-t_k)}-|\delta_k|^2 \leq \frac{K\alpha|\delta_k|^2}{\beta+\alpha}\left(\frac{e^{(\beta+1)(t-t_k)}-1}{\beta+1}-\frac{e^{(1-\alpha)(t-t_k)}-1}{1-\alpha}\right)+\frac{K|P\delta_k|^2}{1-\alpha}\left(e^{(1-\alpha)(t-t_k)}-1\right). \tag{2.3.24}$$

Rearranging the above equation gives (2.3.19). □

2.4 Analysis of Continuous Time 3DVAR

In this section we analyse application of the 3DVAR continuous filtering algorithm for the Lorenz equation (2.2.5). We will use $\{v(t)\}_{t\in[0,\infty)}$ to denote the exact solution sequence from the Lorenz equations which underlies the data; this is to be contrasted with $\{u(t)\}_{t\in[0,\infty)}$ which denotes the random variable which, when conditioned on the data, is approximated by the 3DVAR filter.

We study the continuous time 3DVAR filter, again in the case where $H = (1, 0, 0)$, $\Gamma_0 = \epsilon^2$ and $C = \eta^{-1}\epsilon^2 I$. To analyse the filter it is useful to have the truth v which gives rise to the data appearing in the filter itself. Thus (2.2.17) gives

$$\frac{\mathrm{d}z}{\mathrm{d}t} = Hv + \sqrt{\Gamma_0}\frac{\mathrm{d}w}{\mathrm{d}t}. \tag{2.4.1}$$

We then eliminate z in equation (2.2.19) by using (2.4.1) to obtain

$$\frac{\mathrm{d}m}{\mathrm{d}t} = \mathcal{F}(m) + CH^*\Gamma_0^{-1}H(v-m) + CH^*\Gamma_0^{-\frac{1}{2}}\frac{\mathrm{d}w}{\mathrm{d}t}. \tag{2.4.2}$$

In the specific case of the Lorenz equation we get

$$\frac{\mathrm{d}m}{\mathrm{d}t} = -Am - B(m,m) + f + CH^*\Gamma_0^{-1}H(v-m) + CH^*\Gamma_0^{-\frac{1}{2}}\frac{\mathrm{d}w}{\mathrm{d}t}. \tag{2.4.3}$$

From equation (2.4.2) with the choices of C, H and Γ_0 detailed above we get

$$\frac{\mathrm{d}m}{\mathrm{d}t} = -Am - B(m,m) + f + \frac{1}{\eta}P(v-m) + \frac{\epsilon}{\eta}P\frac{\mathrm{d}w}{\mathrm{d}t} \tag{2.4.4}$$

where we have extended w from a scalar Brownian motion to an $\mathbb{R}^3$-valued Brownian motion for notational convenience. This SDE has a unique global strong solution $m \in C([0,\infty);\mathbb{R}^3)$. Indeed similar techniques used to prove the following result may be used to establish this global existence result, by applying the Itô formula to $|m|^2$ and using the global existence theory in [58]; we omit the details. Recall K given by (2.2.7).

Theorem 2.4.1. *Let m solve equation(2.4.4) and let v solve equation (2.2.5) with initial data $v(0) \in \mathcal{A}$, the global attractor, so that $\sup_{t\geq 0}|v(t)|^2 \leq K$. Then for $\eta K < 4$ we obtain*

$$\mathbb{E}|m(t) - v(t)|^2 \leq e^{-\lambda t}|m(0) - v(0)|^2 + \frac{\epsilon^2}{\eta^2\lambda}(1 - e^{-\lambda t}), \tag{2.4.5}$$

where λ is defined by

$$\lambda = 2\left(1 - \frac{\eta K}{4}\right). \tag{2.4.6}$$

Thus

$$\mathrm{limsup}_{t\to\infty}\mathbb{E}|m(t) - v(t)|^2 \leq \frac{\epsilon^2}{\lambda\eta^2}.$$

Proof. The true solution follows the model

$$\frac{\mathrm{d}v}{\mathrm{d}t} = -Av - B(v,v) + f + \frac{1}{\eta}P(v-v), \tag{2.4.7}$$

where we include the last term, which is identically zero, for clear comparison with the filter equation (2.4.4). Define $\delta = m - v$ and subtract equation (2.4.7) from equation(2.4.4) to obtain

$$\begin{aligned}\frac{\mathrm{d}\delta}{\mathrm{d}t} &= -Am - B(m,m) + Av + B(v,v) - \eta^{-1}P\delta + \epsilon\eta^{-1}P\frac{\mathrm{d}w}{\mathrm{d}t} \\ &= -A\delta - 2B(v,\delta) - B(\delta,\delta) - \eta^{-1}P\delta + \epsilon\eta^{-1}P\frac{\mathrm{d}w}{\mathrm{d}t}.\end{aligned} \tag{2.4.8}$$

Using Itô's formula gives

$$\frac{1}{2}\mathrm{d}|\delta|^2 + \langle A\delta + 2B(v,\delta) + B(\delta,\delta) + \frac{1}{\eta}P\delta, \delta\rangle \mathrm{d}t \leq \langle \epsilon\eta^{-1}P\mathrm{d}w, \delta\rangle + \frac{1}{2}\mathrm{Tr}\left(\epsilon^2\eta^{-2}P\right)\mathrm{d}t. \tag{2.4.9}$$

Using Lemma 2.4.2 and the definition of λ gives

$$\frac{1}{2}\mathrm{d}|\delta|^2 + \frac{\lambda}{2}|\delta|^2\mathrm{d}t \leq \langle \epsilon\eta^{-1}P\mathrm{d}w, \delta\rangle + \frac{1}{2}\mathrm{Tr}\left(\epsilon^2\eta^{-2}P\right)\mathrm{d}t. \tag{2.4.10}$$

Rearranging and taking expectations gives

$$\frac{\mathrm{d}\mathbb{E}|\delta|^2}{\mathrm{d}t} \leq -\lambda\mathbb{E}|\delta|^2 + \frac{\epsilon^2}{\eta^2}. \tag{2.4.11}$$

Use of the Gronwall inequality gives the desired result. □

The following lemma is used in the preceding proof.

Lemma 2.4.2. *Let $v \in \mathcal{A}$. Then*

$$\langle A\delta + 2B(v,\delta) + B(\delta,\delta) + \frac{1}{\eta}P\delta, \delta\rangle \geq \left(1 - \frac{\eta K}{4}\right)|\delta|^2. \tag{2.4.12}$$

Proof. On expanding the inner product

$$\langle A\delta + 2B(v,\delta) + B(\delta,\delta) + \frac{1}{\eta}P\delta, \delta\rangle = \langle A\delta, \delta\rangle + 2\langle B(v,\delta), \delta\rangle + \langle B(\delta,\delta), \delta\rangle + \langle \eta^{-1}P\delta, \delta\rangle. \tag{2.4.13}$$

We now use the Properties 2.2.2(1),(5) and the fact that true solution lies on the global attractor so that $|v| \leq K$. As a consequence we obtain

$$\langle A\delta + 2B(v,\delta) + B(\delta,\delta) + \frac{1}{\eta}P\delta, \delta\rangle \geq |\delta|^2 - K^{\frac{1}{2}}|\delta||P\delta| + \frac{1}{\eta}|P\delta|^2. \tag{2.4.14}$$

Using Young's inequality with parameter θ

$$\langle A\delta + 2B(v,\delta) + B(\delta,\delta) + \frac{1}{\eta}P\delta, \delta\rangle \geq |\delta|^2 - \frac{1}{2\theta}K|P\delta|^2 - \frac{\theta}{2}|\delta|^2 + \frac{1}{\eta}|P\delta|^2. \tag{2.4.15}$$

Taking $\theta = \frac{\eta K}{2}$ yields the desired result

$$\langle A\delta + 2B(v,\delta) + B(\delta,\delta) + \frac{1}{\eta}P\delta, \delta\rangle \geq \left(1 - \frac{\eta K}{4}\right)|\delta|^2. \tag{2.4.16}$$

□

2.5 Numerical Results

In this section we present numerical results illustrating Theorems 2.3.2 and 2.4.1 established in the two preceding sections. All experiments are conducted with the parameters $(\alpha, b, r) = (10, \frac{8}{3}, 28)$. Both the theorems are mean square results. However, some of our numerics are based on a single long-time realization of the filters in question, with time-averaging used in place of ensemble averaging when mean square results are displayed; we highlight when this is done.

2.5.1 Discrete case

Under the assumptions of Theorem 2.3.2 we expect the mean square error in $\delta = |v - m|$ to decrease exponentially until it is of the size of the observational noise squared. Hence we expect the estimate m to converge to a neighbourhood of the true solution v, where the size of the neighbourhood scales as the size of the noise which pollutes in observation, in mean square. The following experiment indicates that similar behaviour is in fact observed pathwise (Figure 2.5.1), as well as in mean square over an ensemble (Figure 2.5.2). We set up the numerical experiments by computing the true solution v of the Lorenz equations using the explicit Euler method, and then adding Gaussian random noise to the observed x-component to create the data. Throughout we fix the parameter $\eta = 0.1$. In Figure 2.5.1 the observational noise ϵ is fixed and in Figure 2.5.2 we vary it over a range of scales.

Figure 2.5.1 concerns the behaviour of a single realization of the filter. Note that the initial error $|v(0) - m(0)|$ is around $\mathbb{E}|v| \approx 10$ and it decays exponentially with time, converging to $\mathcal{O}(\epsilon)$; for this particular case we chose $\epsilon = 1$. A consequence of the second part of Theorem 2.3.2 is that the logarithm of the asymptotic mean squared error $\log \mathbb{E}|\delta|^2$ varies linearly with the logarithm of the standard deviation of noise in the observations (ϵ) and this is illustrated in Figure 2.5.2. To compute the asymptotic mean square error we take two approaches. In the first, for each ϵ, we time-average the error incurred within a single long trajectory of the filter. In the second approach, we consider spatial average over an ensemble of observational noises ν, at a single time after the error has reached equilibrium. In Figure 2.5.2 we observe the log-linear decrease in the asymptotic

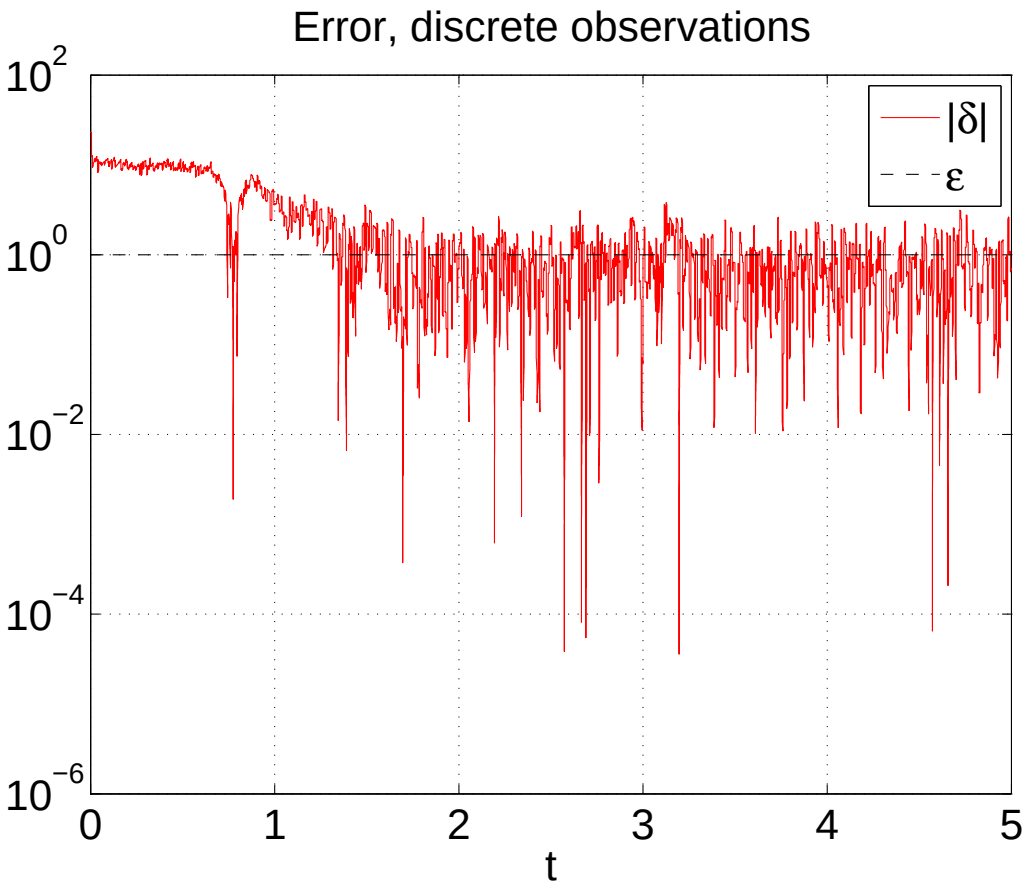


Figure 2.5.1: Decay of initial error from $\mathcal{O}(1)$ to $\mathcal{O}(\epsilon)$ for discrete observations, $\epsilon = 1$, $\eta = 0.1$

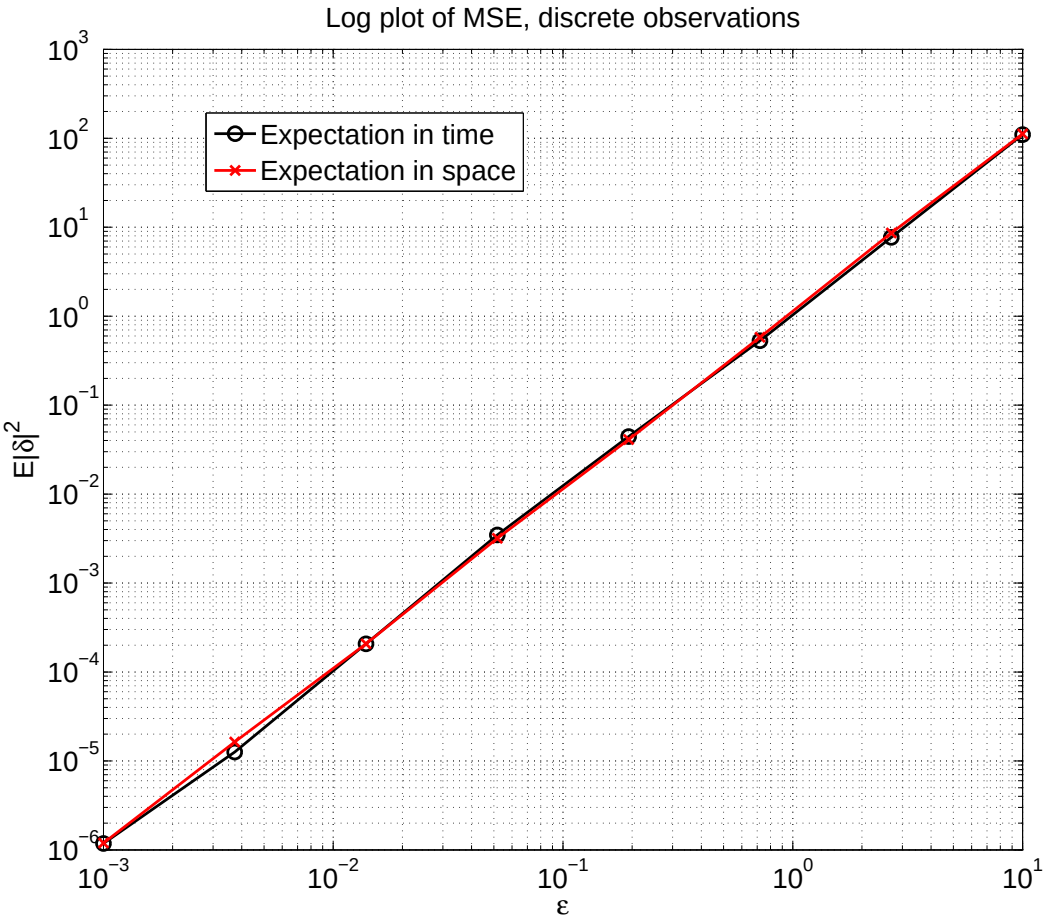


Figure 2.5.2: Log-linear dependence of asymptotic $\mathbb{E}|\delta|^2$ on ϵ for discrete observations, $\eta = 0.1$.

error as the size of the noise decreases; furthermore, the slope of the graph is approximately 2 as predicted by (2.3.10). Both temporal and spatial averaging deliver approximately the same slope.

2.5.2 Continuous case

In the case of continuous observations we again compute a true trajectory of the Lorenz equation using the explicit Euler scheme. We then simulate the SDE (2.4.3) using the Euler-Maruyama method.[1] Similarly to the discrete case, we consider both pathwise and ensemble illustrations of the mean square results in Theorem 2.4.1. Figures 2.5.3 and 2.5.4 concern a single pathwise solution of (2.4.3). Recall from Theorem 2.4.1 that the critical value of η, beneath which the mean square theory holds, is $\eta_c = 4/K$. In Figure 2.5.3 we have $\eta = \frac{1}{2}\eta_c$ whilst in Figure 2.5.4 we have $\eta = 10\eta_c$; in both cases the pathwise error spends most of its time at $\mathcal{O}(\epsilon)$, after the initial transient is removed, suggesting that the critical value of η derived in Theorem 2.4.1 is not sharp. In Figure 2.5.5 we vary the size of observational error ϵ and take $\eta = \frac{1}{8}\eta_c$. The initial error is expected to decay exponentially towards something of order ϵ, and this is what is observed in both the case where averaging is performed in time and in space. Indeed we observe the log-linear decrease in the asymptotic error as the size of the noise decreases, and the slope of the graph is approximately 2, as predicted by equation (2.4.5).

2.6 Conclusions

The study of approximate Gaussian filters for the incorporation of data into high dimensional dynamical systems provides a rich field for applied mathematicians. Potentially such analysis can shed light on algorithms currently in use, whilst also suggesting methods for the improvement of those algorithms. However, rigorous analysis of these filters is in its infancy. The current work demonstrates the properties of the 3DVAR algorithm when applied to the partially observed Lorenz '63 model; it is analogous to the more involved theory developed for the 3DVAR filter applied to the partially observed Navier-Stokes equations in [11, 8]. Work of this type can be built upon in four primary directions: firstly to consider other model dynamical systems of interest to practitioners, such as the Lorenz '96 model [54]; secondly to consider other observation models, such as pointwise velocity field measurements or Lagrangian data for the Navier-Stokes equations, building on the theory of determining modes [34]; thirdly to consider the precise relationships required between the model covariance C and observation operator H to ensure accuracy of the filter; and finally to consider more sophisticated filters such as the extended [32] and ensemble [19, 20] Kalman filters.

We are actively engaged in studying other models, such as Lorenz '96, by similar techniques to those employed here; our work on Lorenz '63 and Navier-Stokes models builds heavily on the synchronization results of Titi and coworkers and we believe that generalization of synchronization

[1] Note that this is equivalent to creating the data z from (2.4.1) and solving (2.2.19) and, since we have access to the truth, is computationally expedient.

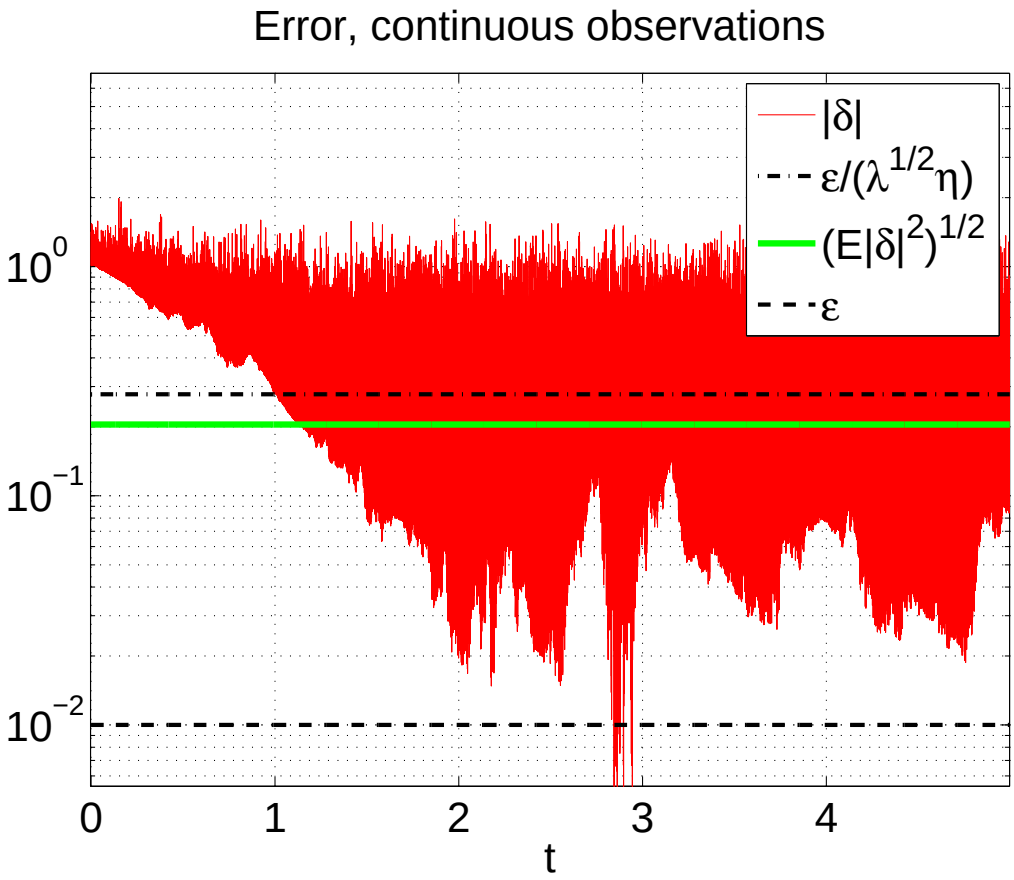


Figure 2.5.3: Decay of initial error from $\mathcal{O}(1)$ to $\mathcal{O}(\epsilon)$ for continuous observations, $\epsilon = 0.01$. Results are shown for $\eta = 2/K < \eta_c$.

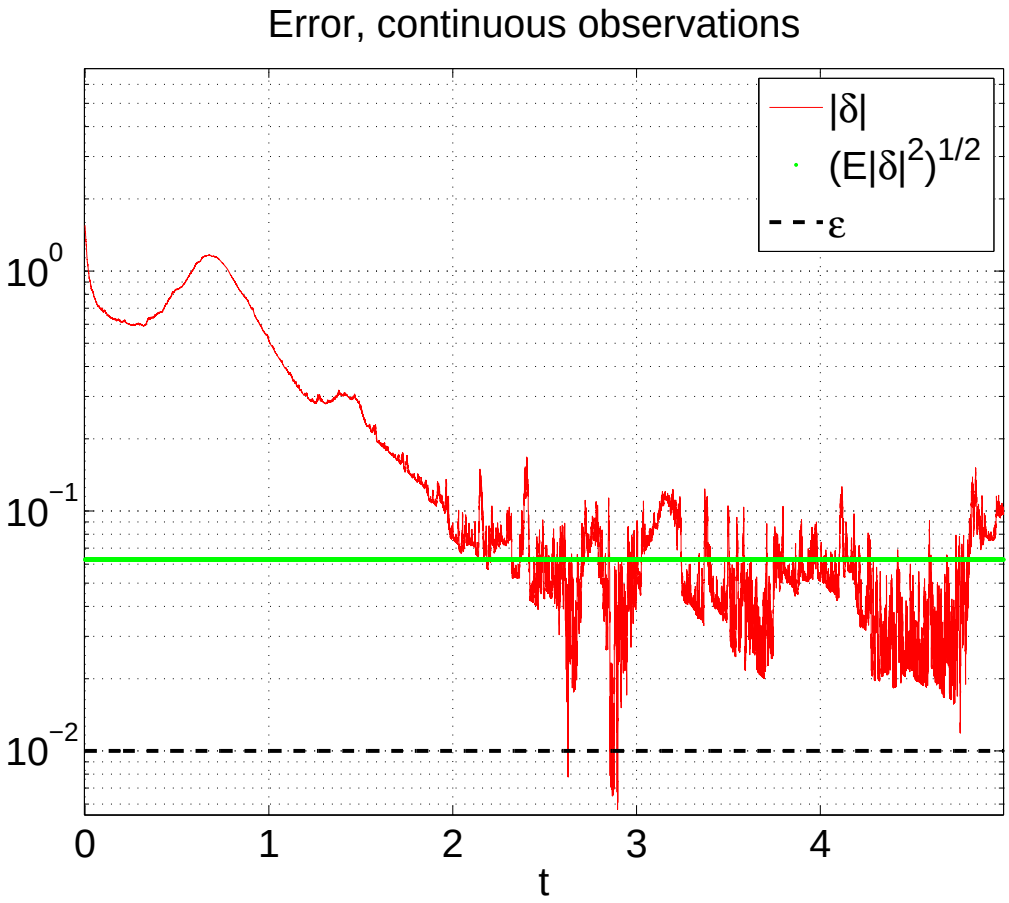


Figure 2.5.4: Decay of initial error from $\mathcal{O}(1)$ to $\mathcal{O}(\epsilon)$ for continuous observations, $\epsilon = 0.01$. Results are shown for $\eta = 40/K = 10\eta_c$.

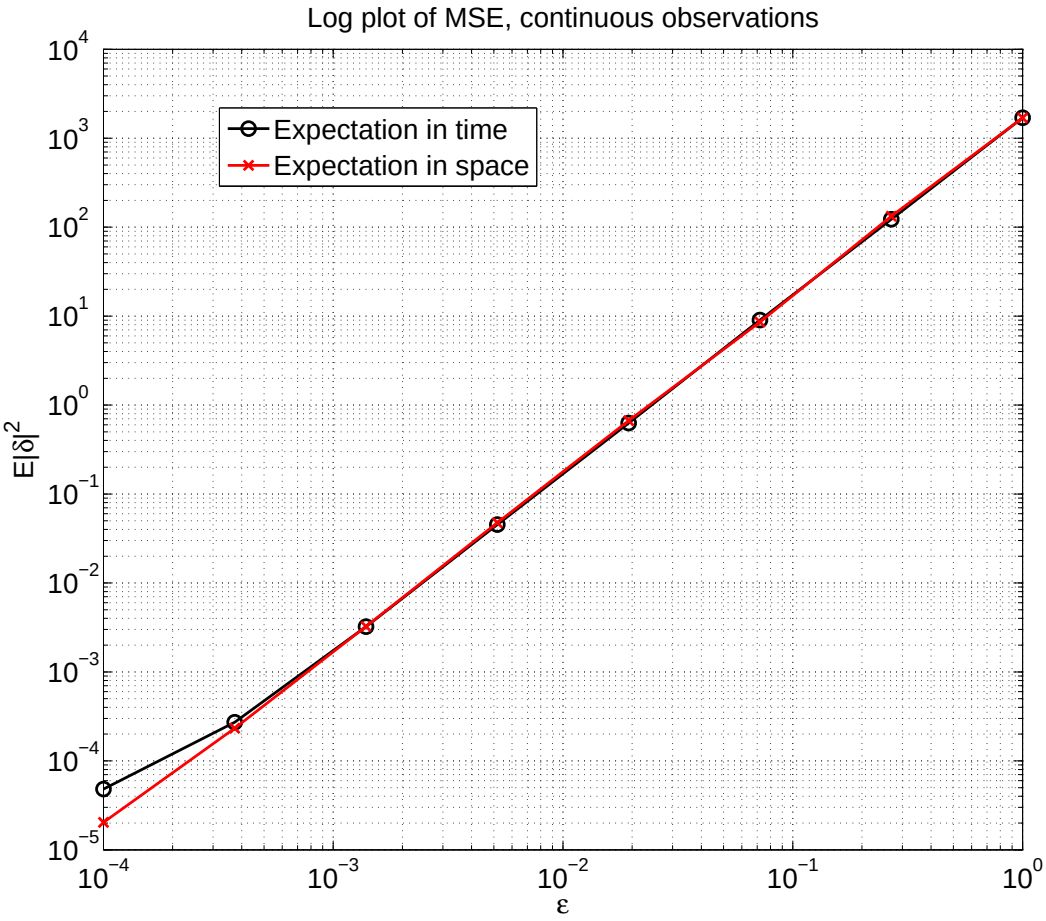


Figure 2.5.5: Log-linear dependence of asymptotic $\log \mathbb{E}|\delta|^2$ on $\log \epsilon$ for continuous observations and $\eta = 1/(2K)$.

properties is a key first step in the study of other models. Regarding the second direction, Lagrangian data introduces an additional auxiliary system for the observed variables through which the system of interest is observed, necessitating careful design of correlations in the design parameters C, meaning that the analysis will be considerably more complicated than for Eulerian data. This links to the third direction: in general the relationship between the model covariance and observation operator required to obtain filter accuracy may be quite complicated and is an important avenue for study in this field; even for the particular Lorenz '63 model studied herein, with observation of only the x component of the system, this complexity is manifest if the covariance is not diagonal. Relating to the fourth and final direction, it is worth noting that 3DVAR is outdated operationally and empirical studies of filter accuracy have recently been focused on the more sophisticated methods such as ensemble Kalman filter and 4DVAR [37, 42]. These empirical studies indicate that the more sophisticated methods outperform 3DVAR, as expected, and therefore suggest the importance of rigorous analysis of those methods.

Chapter 3

Partial observations on Lorenz'96 system

3.1 Introduction

Data assimilation is concerned with the blending of data and dynamical mathematical models, often in an online fashion where it is known as filtering; motivation comes from applications in the geophysical sciences such as weather forecasting [36], oceanography [6] and oil reservoir simulation [59]. Over the last decade there has been a growing body of theoretical understanding which enables use of the theory of synchronization in dynamical systems to establish desirable properties of these filters. This idea is highlighted in the recent book [1] from a physics perspective and, on the rigorous mathematical side, has been developed from a pair of papers by Olson, Titi and co-workers [60, 29], in the context of the Navier-Stokes equation in which a finite number of Fourier modes are observed. This mathematical work of Olson and Titi concerns perfect (noise-free) observations, but the ideas have been extended to the incorporation of noisy data for the Navier-Stokes equation in the papers [8, 11]. Furthermore the techniques used are quite robust to different dissipative dynamical systems, and have been demonstrated to apply in the Lorenz '63 model [29, 41], and also to point-wise in space and continuous time observations [4] by use of a control theory perspective similar to that which arises from the derivation of continuous time limits of discrete time filters [8]. A key question in the field is to determine relationships between the underlying dynamical system and the observation operator which are sufficient to ensure that the signal can be accurately recovered from a chaotic dynamical system, whose initialization is not known precisely, by the use of observed data. Our purpose is to investigate this question theoretically and computationally. We work in the context of the Lorenz '96 model, widely adopted as a useful test model in the atmospheric sciences data assimilation community [56, 61].

The primary contributions of the paper are: (i) to theoretically demonstrate the robustness of the methodology proposed by Olson and Titi, by extending it to the Lorenz '96 model; (ii) to highlight the gap between such theories and what can be achieved in practice, by performing

careful numerical experiments; and (iii) to illustrate the power of allowing the observation operator to adapt to the dynamics as this leads to accurate reconstruction of the signal based on very sparse observations. Indeed our approach in (iii) suggests highly efficient new algorithms where the observation operator is allowed to adapt to the current state of the dynamical system. The question of how to optimize the observation operator to maximize information was first addressed in the context of atmospheric science applications in [55]. The adaptive observation operators that we propose are not currently practical for operational atmospheric data assimilation, but they suggest a key principle which should underlie the construction of adaptive observation operators: to learn as much as possible about modes of instability in the dynamics at minimal cost.

The outline of the paper is as follows. In section 3.2 we introduce the model set up and a family of Kalman-based filtering schemes which include as particular cases the Three-dimensional Variational method (3DVAR) and the Extended Kalman Filter (ExKF) used in this paper. All of these methods may be derived from sequential application of a minimization principle which encodes the trade-off between matching the model and matching the data. In section 3.3 we describe the Lorenz '96 model and discuss its properties that are relevant to this work. In section 3.4 we introduce a fixed observation operator which corresponds to observing two thirds of the signal and study theoretical properties of the 3DVAR filter, in both a continuous and a discrete time setting. In section 3.5 we introduce an adaptive observation operator which employs knowledge of the linearized dynamics over the assimilation window to ensure that the unstable directions of the dynamics are observed. We then numerically study the performance of a range of filters using the adaptive observations. In subsection 3.5.1 we consider the 3DVAR method, whilst subsection 3.5.2 focuses on the Extended Kalman Filter (ExKF). In subsection 3.5.2 we also compare the adaptive observation implementation of the ExKF with the AUS scheme [78] which motivates our work. The AUS scheme projects the model covariances into the subspaces governed by the unstable dynamics, whereas we use this idea on the observation operators themselves, rather than on the covariances. In section 3.6 we summarize the work and draw some brief conclusions. In order to maintain a readable flow of ideas, the proofs of all properties, propositions and theorems stated in the main body of the text are collected in an appendix.

Throughout the paper we denote by $\langle \cdot, \cdot \rangle$ and $|\cdot|$ the standard Euclidean inner-product and norm. For positive-definite matrix C we define $|\cdot|_C := |C^{-\frac{1}{2}} \cdot|$.

3.2 Set Up

We consider the ordinary differential equation (ODE)

$$\frac{\mathrm{d}v}{\mathrm{d}t} = \mathcal{F}(v), \quad v(0) = v_0, \tag{3.2.1}$$

where the solution to (3.2.1) is referred to as the *signal*. We denote by $\Psi : \mathbb{R}^J \times \mathbb{R}^+ \to \mathbb{R}^J$ the solution operator for the equation (3.2.1), so that $v(t) = \Psi(v_0; t)$. In our discrete time filtering

developments we assume that, for some fixed $h > 0$, the signal is subject to observations at times $t_k := kh$, $k \geq 1$. We then write $\Psi(\cdot) := \Psi(\cdot; h)$ and $v_k := v(kh)$, with slight abuse of notation to simplify the presentation. Our main interest is in using partial observations of the discrete time dynamical system

$$v_{k+1} = \Psi(v_k), \quad k \geq 0, \tag{3.2.2}$$

to make estimates of the state of the system. To this end we introduce the family of linear observation operators $\{H_k\}_{k\geq 1}$, where $H_k : \mathbb{R}^J \to \mathbb{R}^M \leq \mathbb{R}^J$ is assumed to have rank (which may change with k) less than or equal to $M \leq J$. We then consider data $\{y_k\}_{k\geq 1}$ given by

$$y_k = H_k v_k + \nu_k, \quad k \geq 1, \tag{3.2.3}$$

where we assume that the random and/or systematic error ν_k (and hence also y_k) is contained in $H_k\mathbb{R}^J$. If $Y_k = \{y_\ell\}_{\ell=1}^k$ then the objective of filtering is to estimate v_k from Y_k given incomplete knowledge of v_0; furthermore this is to be done in a sequential fashion, using the estimate of v_k from Y_k to determine the estimate of v_{k+1} from Y_{k+1}. We are most interested in the case where $M < J$, so that the observations are partial, and $H_k\mathbb{R}^J$ is a strict subset of $\mathbb{R}^J$; in particular we address the question of how small M can be chosen whilst still allowing accurate recovery of the signal over long time-intervals.

Let m_k denote our estimate of v_k given Y_k. The discrete time filters used in this paper have the form

$$m_{k+1} = \text{argmin}_m \left\{ \frac{1}{2}\left|m - \Psi(m_k)\right|^2_{\widehat{C}_{k+1}} + \frac{1}{2}\left|y_{k+1} - H_{k+1}m\right|^2_{\Gamma} \right\}. \tag{3.2.4}$$

The norm in the second term is only applied within the M-dimensional image space of H_{k+1}, where y_{k+1} lies; then Γ is realized as a positive-definite $M \times M$ matrix in this image space, and $\widehat{C}_{k+1}$ is a positive-definite $J \times J$ matrix. The minimization represents a compromise between respecting the model and respecting the data, with the covariance weights $\widehat{C}_{k+1}$ and Γ determining the relative size of the two contributions; see [43] for more details. Different choices of $\widehat{C}_{k+1}$ give different filtering methods. For instance, the choice $\widehat{C}_{k+1} = C_0$ (constant in k) corresponds to the 3DVAR method. More sophisticated algorithms, such as the ExKF, allow $\widehat{C}_{k+1}$ to depend on m_k.

All the discrete time algorithms we consider proceed iteratively in the sense that the estimate m_{k+1} is determined by the previous one, m_k, and the observed data y_{k+1}; we are given an initial condition m_0 which is an imperfect estimate of v_0. It is convenient to see the update $m_k \mapsto m_{k+1}$ as a two-step process. In the first one, known as the *forecast step*, the estimate m_k is evolved with the dynamics of the underlying model yielding a prediction $\Psi(m_k)$ for the current state of the system. In the second step, known as the *analysis step*, the forecast is used in conjunction with the observed data y_{k+1} to produce the estimate m_{k+1} of the true state of the underlying system v_{k+1}, using the minimization principle (3.2.4).

In section 3.4 we study the continuous time filtering problem for fixed observation operator,

where the goal is to estimate the value of a continuous time signal

$$v(t) = \Psi(v_0, t), \quad t \geq 0,$$

at time $T > 0$. As in the discrete case, it is assumed that only incomplete knowledge of v_0 is available. In order to estimate $v(T)$ we assume that we have access, at each time $0 < t \leq T$, to a (perhaps noisily perturbed) projection of the signal given by a fixed, constant in time, observation matrix H. The continuous time limit of 3DVAR with constant observation operator H, is obtained by setting $\Gamma = h^{-1}\Gamma_0$ and $\widehat{C}_{k+1} = C$ and letting $h \to 0$. The resulting filter, derived in [8], is given by

$$\frac{\mathrm{d}m}{\mathrm{d}t} = \mathcal{F}(m) + CH^*\Gamma_0^{-1}\Big(\frac{\mathrm{d}z}{\mathrm{d}t} - Hm\Big), \tag{3.2.5}$$

where the observed data is now z – formally the time-integral of the natural continuous time limit of y – which satisfies the stochastic differential equation (SDE)

$$\frac{\mathrm{d}z}{\mathrm{d}t} = Hv + H\Gamma_0^{\frac{1}{2}}\frac{\mathrm{d}w}{\mathrm{d}t}, \tag{3.2.6}$$

for w a unit Wiener process. This filter has the effect of nudging the solution towards the observed data in the H-projected direction. A similar idea is used in [4] to assimilate pointwise observations of the Navier-Stokes equation.

For the discrete and continuous time filtering schemes as described we address the following questions:

- how does the filter error $|m_k - v_k|$ behave as $k \to \infty$ (discrete setting)?
- how does the filter error $|m(t) - v(t)|$ behave as $t \to \infty$ (continuous setting)?

We answer these questions in the section 3.4 in the context of the Lorenz '96 model: for a carefully chosen fixed observation operator we determine conditions under which the large time filter error is small – this is filter accuracy. We then turn to the adaptive observation operator and focus on the following lines of enquiry:

- how much do we need to observe to obtain filter accuracy? (in other words what is the minimum rank of the observation operator required?)
- how does adapting the observation operator affect the answer to this question?

We study both these questions numerically in section 3.5, again focussing on the Lorenz '96 model to illustrate ideas.

3.3 Lorenz '96 Model

The Lorenz '96 model is a lattice-periodic system of coupled nonlinear ODE whose solution $u = (u^{(1)}, \ldots, u^{(J)})^T \in \mathbb{R}^J$ satisfies

$$\frac{\mathrm{d}u^{(j)}}{\mathrm{d}t} = u^{(j-1)}(u^{(j+1)} - u^{(j-2)}) - u^{(j)} + F \quad \text{for } j = 1, 2, \cdots, J, \tag{3.3.1}$$

subject to the periodic boundary conditions

$$u^{(0)} = u^{(J)}, \qquad u^{(J+1)} = u^{(1)}, \qquad u^{(-1)} = u^{(J-1)}. \tag{3.3.2}$$

Here F is a forcing parameter, constant in time. For our numerical experiments we will choose F so that the dynamical system exhibits sensitive dependence on initial conditions and positive Lyapunov exponents. For example, for $F = 8$ and $J = 60$ the system is chaotic. Our theoretical results apply to any choice of the parameter F and to arbitrarily large system dimension J.

It is helpful to write the model in the following form, widely adopted in the analysis of geophysical models as dissipative dynamical systems [74]:

$$\frac{\mathrm{d}u}{\mathrm{d}t} + Au + B(u, u) = f, \quad u(0) = u_0 \tag{3.3.3}$$

where

$$A = I_{J\times J}, \quad f = \begin{pmatrix} F \\ \vdots \\ F \end{pmatrix}_{J\times 1}$$

and for $u, \tilde{u} \in \mathbb{R}^J$

$$B(u, \tilde{u}) = -\frac{1}{2}\begin{pmatrix} \tilde{u}^{(2)}u^{(J)} + u^{(2)}\tilde{u}^{(J)} - \tilde{u}^{(J)}u^{(J-1)} - u^{(J)}\tilde{u}^{(J-1)} \\ \vdots \\ \tilde{u}^{(j-1)}u^{(j+1)} + u^{(j-1)}\tilde{u}^{(j+1)} - \tilde{u}^{(j-2)}u^{(j-1)} - u^{(j-2)}\tilde{u}^{(j-1)} \\ \vdots \\ \tilde{u}^{(J-1)}u^{(1)} + u^{(J-1)}\tilde{u}^{(1)} - \tilde{u}^{(J-2)}u^{(J-1)} - u^{(J-2)}\tilde{u}^{(J-1)} \end{pmatrix}_{J\times 1}.$$

We will use the following properties of A and B, proved in the Appendix:

Properties 3.3.1. *For $u, \tilde{u} \in \mathbb{R}^J$*

1. $\langle Au, u\rangle = |u|^2$.
2. $\langle B(u, u), u\rangle = 0$.
3. $B(u, \tilde{u}) = B(\tilde{u}, u)$.

4. $|B(u,\tilde{u})| \leq 2|u||\tilde{u}|$.

5. $2\langle B(u,\tilde{u}),u\rangle = -\langle B(u,u),\tilde{u}\rangle$.

Property (1) shows that the linear term induces dissipation in the model, whilst property (2) shows that the nonlinear term is energy-conserving. Balancing these two properties against the injection of energy through f gives the existence of an absorbing, forward-invariant ball for equation (3.3.3), as stated in the following proposition, proved in the Appendix.

Proposition 3.3.2. *Let $K = 2JF^2$ and define $\mathcal{B} := \{u \in \mathbb{R}^J : |u|^2 \leq K\}$. Then $\mathcal{B}$ is an absorbing, forward-invariant ball for equation* (3.3.3)*: for any $u_0 \in \mathbb{R}^J$ there is time $T = T(|u_0|) \geq 0$ such that $u(t) \in \mathcal{B}$ for all $t \geq T$.*

3.4 Fixed Observation Operator

In this section we consider filtering the Lorenz '96 model with a specific choice of fixed observation matrix P (thus $H_k = H = P$) that we now introduce. First, we let $\{e_j\}_{j=1}^J$ be the standard basis for the Euclidean space $\mathbb{R}^J$ and assume that $J = 3J'$ for some $J' \geq 1$. Then the projection matrix P is defined by replacing every third column of the identity matrix $I_{J\times J}$ by the zero vector:

$$P = \Big(\; e_1,\, e_2,\, 0,\, e_4,\, e_5,\, 0, \ldots \; \Big)_{J\times J}. \tag{3.4.1}$$

Thus P has rank $M = 2J'$. We also define its complement Q as

$$Q = I_{J\times J} - P.$$

Remark 3.4.1. *Note that in the definition of the projection matrix P we could have chosen either the first or the second column to be set to zero periodically, instead of choosing every third column this way; the theoretical results in the the remainder of this section would be unaltered by doing this.*

The matrix P provides sufficiently rich observations to allow the accurate recovery of the signal in the long-time asymptotic regime, both in continuous and discrete time settings. The following property of P, proved in the appendix, plays a key role in the analysis:

Properties 3.4.2. *The bilinear form $B(\cdot,\cdot)$ as defined after* (3.3.3) *satisfies $B(Qu,Qu) = 0$ and, furthermore, there is a constant $c > 0$ such that*

$$|\langle B(u,u),\tilde{u}\rangle| \leq c|u||\tilde{u}||Pu|.$$

All proofs in the following subsections are given in the Appendix.

3.4.1 Continuous Assimilation

In this subsection we assume that the data arrives continuously in time. Subsection 3.4.1 deals with noiseless data, and the more realistic noisy scenario is studied in subsection 3.4.1. We aim to show that, in the large time asymptotic, the filter is close to the truth. In the absence of noise our results are analogous to those for the partially observed Lorenz '63 and Navier-Stokes models in [60]; in the presence of noise the results are similar to those proved in [8] for the Navier-Stokes equation and in [41] for the Lorenz '63 model, and generalize the work in [73] to non-globally Lipschitz vector fields.

Noiseless Observations

The true solution v satisfies the following equation

$$\frac{\mathrm{d}v}{\mathrm{d}t} + v + B(v,v) = f, \quad v(0) = v_0. \tag{3.4.2}$$

Suppose that the projection Pv of the true solution is perfectly observed and continuously assimilated into the approximate solution m. The *synchronization filter* m has the following form:

$$m = Pv + q, \tag{3.4.3}$$

where v is the true solution given by (3.4.2) and q satisfies the equation (3.3.3) projected by Q to obtain

$$\frac{\mathrm{d}q}{\mathrm{d}t} + q + QB(Pv+q, Pv+q) = Qf, \quad q(0) = q_0. \tag{3.4.4}$$

Equations (3.4.3) and (3.4.4) form the continuous time synchronization filter. The following theorem shows that the approximate solution converges to the true solution asymptotically as $t \to \infty$.

Theorem 3.4.3. *Let m be given by the equations* (3.4.3), (3.4.4) *and let v be the solution of the equation* (3.4.2) *with initial data $v_0 \in \mathcal{B}$, the absorbing ball in Proposition 3.3.2, so that $\sup_{t\geq 0}|v(t)|^2 \leq K$. Then*

$$\lim_{t\to\infty} |m(t) - v(t)|^2 = 0.$$

The result establishes that in the case of high frequency in time observations the approximate solution converges to the true solution even though the signal is observed partially at frequency 2/3 in space. We now extend this result by allowing for noisy observations.

Noisy Observations: Continuous 3DVAR

Recall that the continuous time limit of 3DVAR is given by (3.2.5) where the observed data z, the integral of y, satisfies the SDE (3.2.6). We study this filter in the case where $H = P$ and under small observation noise $\Gamma_0 = \epsilon^2 I$. The 3DVAR model covariance is then taken to be of the size

of the observation noise. We choose $C = \sigma^2 I$, where $\sigma^2 = \sigma^2(\epsilon) = \eta^{-1}\epsilon^2$, for some $\eta > 0$. Then equations (3.2.5) and (3.2.6) can be rewritten as

$$\frac{\mathrm{d}m}{\mathrm{d}t} = \mathcal{F}(m) + \frac{1}{\eta}\Big(\frac{\mathrm{d}z}{\mathrm{d}t} - Pm\Big) \tag{3.4.5}$$

where

$$\frac{\mathrm{d}z}{\mathrm{d}t} = Pv + \epsilon P\frac{\mathrm{d}w}{\mathrm{d}t}, \tag{3.4.6}$$

and w is a unit Wiener process. Note that the parameter ϵ represents both the size of the 3DVAR observation covariance and the size of the noise in the observations.

The reader will notice that the continuous time synchronization filter is obtained from this continuous time 3DVAR filter if ϵ is set to zero and if the (singular) limit $\eta \to 0$ is taken. The next theorem shows that the approximate solution m converges to a neighbourhood of the true solution v where the size of the neighbourhood depends upon ϵ. Similarly as in [41] and [8] it is required that η, the ratio between the size of observation and model covariances, is sufficiently small. The next theorem is thus a natural generalization of Theorem 3.4.3 to incorporate noisy data.

Theorem 3.4.4. *Let (m, z) solve the equations (3.4.5), (3.4.6) and let v solve the equation (3.4.2) with the initial data $v(0) \in \mathcal{B}$, the absorbing ball of Proposition 3.3.2, so that $\sup_{t\geq 0}|v(t)|^2 \leq K$. Then for the constant c as given in the Property 3.4.2, given $\eta < \frac{4}{c^2 K}$ we obtain*

$$\mathbb{E}|m(t) - v(t)|^2 \leq e^{-\lambda t}|m(0) - v(0)|^2 + \frac{2J\epsilon^2}{3\lambda\eta^2}(1 - e^{-\lambda t}), \tag{3.4.7}$$

where λ is defined by

$$\lambda = 2\left(1 - \frac{c^2\eta K}{4}\right). \tag{3.4.8}$$

Thus

$$\mathrm{limsup}_{t\to\infty}\mathbb{E}|m(t) - v(t)|^2 \leq a\epsilon^2,$$

where $a = \frac{2J}{3\lambda\eta^2}$ does not depend on the strength of the observation noise, ϵ.

3.4.2 Discrete Assimilation

We now turn to discrete data assimilation. Recall that filters in discrete time can be split into two steps: forecast and analysis. In this section we establish conditions under which the corrections made at the analysis steps overcome the divergence inherent due to nonlinear instabilities of the model in the forecast stage. As in the previous section we study first the case of noiseless data, generalizing the work of [29] from the Navier-Stokes and Lorenz '63 models to include the Lorenz '96 model, and then study the case of 3DVAR, generalizing the work in [11, 41], which concerns the Navier-Stokes and Lorenz '63 models respectively, to the Lorenz '96 model.

Noiseless Observations

Let $h > 0$, and set $t_k := kh$, $k \geq 0$. For any function $g : \mathbb{R}^+ \to \mathbb{R}^J$, continuous in $[t_{k-1}, t_k)$, we denote $g(t_k^-) := \lim_{t \uparrow t_k} g(t)$. Let v be a solution of equation (3.4.2) with $v(0)$ in the absorbing forward-invariant ball $\mathcal{B}$. The discrete time synchronization filter m of [29] may be expressed as follows:

$$\frac{\mathrm{d}m}{\mathrm{d}t} + m + B(m, m) = f, \quad t \in (t_k, t_{k+1}), \tag{3.4.9a}$$

$$m(t_k) = Pv(t_k) + Qm(t_k^-). \tag{3.4.9b}$$

Thus the filter consists of solving the underlying dynamical model, by resetting the filter to take the value $Pv(t)$ in the subspace $P\mathbb{R}^J$ at every time $t = t_k$. The following theorem shows that the filter m converges to the true signal v.

Theorem 3.4.5. *Let v be a solution of the equation (3.4.2) with $v(0) \in \mathcal{B}$. Then there exists $h^* > 0$ such that for any $h \in (0, h^*]$ the approximating solution m given by (3.4.9) converges to v as $t \to \infty$.*

Noisy Observations: Discrete 3DVAR

Now we consider the situation where the data is noisy and $H_k = P$. We employ the 3DVAR filter which results from the minimization principle (3.2.4) in the case where $\widehat{C}_{k+1} = \sigma^2 I$ and $\Gamma = \epsilon^2 I$. Recall the true signal is determined by the equation (3.2.2) and the observed data by the equation (3.2.3), now written in terms of the true signal $v_k = v(t_k)$ solving the equation (3.3.3) with $v_0 \in \mathcal{B}$. Thus

$$\begin{aligned} v_{k+1} &= \Psi(v_k), \quad v_0 \in \mathcal{B}, \\ y_{k+1} &= Pv_{k+1} + \nu_{k+1}. \end{aligned}$$

If we define $\eta := \frac{\epsilon^2}{\sigma^2}$ then the 3DVAR filter can be written as

$$m_{k+1} = \Big(\frac{\eta}{1+\eta} P + Q\Big)\Psi(m_k) + \frac{1}{1+\eta} y_{k+1},$$

after noting that $Py_{k+1} = y_{k+1}$ because P is a projection and ν_{k+1} is assumed to lie in the image of P. In fact the data has the following form:

$$\begin{aligned} y_{k+1} &= Pv_{k+1} + P\nu_{k+1} \\ &= P\Psi(v_k) + \nu_{k+1}. \end{aligned}$$

Combining the two equations gives

$$m_{k+1} = \Big(\frac{\eta}{1+\eta} P + Q\Big)\Psi(m_k) + \frac{1}{1+\eta}\Big(P\Psi(v_k) + \nu_{k+1}\Big). \tag{3.4.10}$$

We can write the equation for the true solution v_k, given by (3.2.2), in the following form:

$$v_{k+1} = \Big(\frac{\eta}{1+\eta}P + Q\Big)\Psi(v_k) + \frac{1}{1+\eta}P\Psi(v_k). \tag{3.4.11}$$

Note that $v_k = v(t_k)$ where $v(\cdot)$ solves (3.4.2). We are interested in comparing the output of the filter, m_k, with the true signal v_k. Notice that if the noise ν_k is set to zero and if the limit $\eta \to 0$ is taken then the filter becomes

$$m_{k+1} = P\Psi(v_k) + Q\Psi(m_k)$$

which is precisely the discrete time synchronization filter. Theorem 3.4.6 below will reflect this observation, constituting a noisy variation on Theorem 3.4.5.

We will assume that the ν_k are independent random variables that satisfy the bound $|\nu_k| \le \epsilon$, thereby linking the scale of the covariance Γ employed in 3DVAR to the size of the noise. We let $\|\cdot\|$ be the norm defined by $\|z\| := |z| + |Pz|$, $z \in \mathbb{R}^J$.

Theorem 3.4.6. *Let v be the solution of the equation (3.4.2) with $v(0) \in \mathcal{B}$. Assume that $\{\nu_k\}_{k\ge1}$ is a sequence of independent bounded random variables such that, for every k, $|\nu_k| \le \epsilon$. Then there are choices (detailed in the proof in the appendix) of assimilation step $h > 0$ and parameter $\eta > 0$ sufficiently small such that, for some $\alpha \in (0,1)$ and provided that the noise $\epsilon > 0$ is small enough, the error satisfies*

$$\|m_{k+1} - v_{k+1}\| \le \alpha\|m_k - v_k\| + 2\epsilon. \tag{3.4.12}$$

Thus, there is $a > 0$ such that

$$\limsup_{k\to\infty} \|m_k - v_k\| \le a\epsilon.$$

3.5 Adaptive Observation Operator

The theory in the previous section demonstrates that accurate filtering of chaotic models is driven by observing enough of the dynamics to control the exponential separation of trajectories in the dynamics. However the fixed observation operator P that we analyze requires observation of 2/3 of the system state vector. Even if the observation operator is fixed our numerical results will show that observation of this proportion of the state is not necessary to obtain accurate filtering. Furthermore, by adapting the observations to the dynamics, we will be able to obtain the same quality of reconstruction with even fewer observations. In this section we will demonstrate these ideas in the context of noisy discrete time filtering, and with reference to the Lorenz '96 model.

The variational equation for the dynamical system (3.2.1) is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}D\Psi(u,t) = D\mathcal{F}\big(\Psi(u,t)\big)\cdot D\Psi(u,t); \qquad D\Psi(u,0) = I_{J\times J}, \tag{3.5.1}$$

using the chain rule. The solution of the variational equation gives the derivative matrix of the solution operator Ψ, which in turn characterizes the behaviour of Ψ with respect to small variations in the initial value u. Let $L_{k+1} := L(t_{k+1})$ be the solution of the variational equation (3.5.1) over the assimilation window (t_k, t_{k+1}), initialized at $I_{J\times J}$, given as

$$\frac{\mathrm{d}L}{\mathrm{d}t} = D\mathcal{F}\big(\Psi(m_k, t-t_k)\big)L, \quad t \in (t_k, t_{k+1}); \qquad L(t_k) = I_{J\times J}. \tag{3.5.2}$$

Let $\{\lambda_k^j, \psi_k^j\}_{j=1}^J$ denote eigenvalue/eigenvector pairs of the matrix $L_{k+1}^T L_{k+1}$, where the eigenvalues (which are, of course, real) are ordered to be non-decreasing, and the eigenvectors are orthonormalized with respect to the Euclidean inner-product $\langle\cdot,\cdot\rangle$. We define the adaptive observation operator H_k to be

$$H_k := H_0(\psi_k^1, \cdots, \psi_k^J)^T \tag{3.5.3}$$

where

$$H_0 = \begin{pmatrix} 0 & 0 \\ 0 & I_{M\times M} \end{pmatrix}. \tag{3.5.4}$$

Thus H_0 and H_k both have rank M. Defined in this way we see that for any given $v \in \mathbb{R}^J$ the projection $H_k v$ is given by the vector

$$\Big(0, \cdots, 0, \langle\psi_k^{J-M+1}, v\rangle, \cdots, \langle\psi_k^J, v\rangle\Big)^T,$$

that is the projection of v onto the M eigenvectors of $L_{k+1}^T L_{k+1}$ with largest modulus.

Remark 3.5.1. *In the following work we consider the leading eigenvalues and corresponding eigenvectors of the matrix $L_k^T L_k$ to track the unstable (positive Lyapunov growth) directions. To leading order in h it is equivalent to consider the matrix $L_k L_k^T$ in the case of frequent observations (small h) as can be seen by the following expressions*

$$\begin{aligned} L_k^T L_k &= (I + hD\mathcal{F}_k)^T(I + hD\mathcal{F}_k) + \mathcal{O}(h^2) \\ &= I + h(D\mathcal{F}_k^T + D\mathcal{F}_k) + \mathcal{O}(h^2) \end{aligned}$$

and

$$\begin{aligned} L_k L_k^T &= (I + hD\mathcal{F}_k)(I + hD\mathcal{F}_k)^T + \mathcal{O}(h^2) \\ &= I + h(D\mathcal{F}_k + D\mathcal{F}_k^T) + \mathcal{O}(h^2), \end{aligned}$$

where $D\mathcal{F}_k = D\mathcal{F}(m_k)$.

Of course for large intervals h, the above does not hold, and the difference between $L_k^T L_k$ and $L_k L_k^T$ may be substantial. It is however clear that these operators have the same eigenvalues, with the eigenvectors of $L_k L_k^T$ corresponding to λ_k^j given by $L_k\psi_k^j$ for the corresponding eigenvector

ψ_k^j of $L_k^T L_k$. That is to say, for the linearized deformation map L_k, the direction ψ_k^j is the pre-deformation principle direction corresponding to the principle strain λ_k^j induced by the deformation. The direction $L_k\psi_k^j$ is the post-deformation principle direction corresponding to the principle strain λ_k^j. The dominant directions chosen in Eq. (3.5.3) *are those directions corresponding to the greatest growth over the interval (t_k, t_{k+1}) of infinitesimal perturbations to the predicting trajectory, $\Psi(m_{k-1}, h)$ at time t_k. This is only one sensible option. One could alternatively consider the directions corresponding to the greatest growth over the interval (t_{k-1}, t_k), or over the whole interval (t_{k-1}, t_{k+1}). Investigation of these alternatives is beyond the scope of this work and is therefore deferred to later investigation.*

We make a small shift of notation and now consider the observation operator H_k as a linear mapping from $\mathbb{R}^J$ into $\mathbb{R}^M$, rather than as a linear operator from $\mathbb{R}^J$ into itself, with rank M; the latter perspective was advantageous for the presentation of the analysis, but differs from the former which is sometimes computationally advantageous and more widely used for the description of algorithms. Recall the minimization principle (3.2.4), noting that now the first norm is in $\mathbb{R}^J$ and the second in $\mathbb{R}^M$.

3.5.1 3DVAR

Here we consider the minimization principle (3.2.4) with the choice $\widehat{C}_{k+1} = C_0 \in \mathbb{R}^{J\times J}$, a strictly positive-definite matrix, for all k. Assuming that $\Gamma \in \mathbb{R}^{M\times M}$ is also strictly positive-definite, the filter may be written as

$$m_{k+1} = \Psi(m_k) + G_{k+1}\Big(y_{k+1} - H_{k+1}\Psi(m_k)\Big) \tag{3.5.5a}$$

$$G_{k+1} = C_0 H_{k+1}^T (H_{k+1} C_0 H_{k+1}^T + \Gamma)^{-1}. \tag{3.5.5b}$$

As well as using the choice of H_k defined in (3.5.3), we also employ the fixed observation operator where $H_k = H$, including the choice $H = P$ given by (3.4.1). In the last case $J = 3J'$, $M = 2J'$ and P is realized as a$2J' \times 3J'$ matrix.

We make the choices $C_0 = \sigma^2 I_{J\times J}$, $\Gamma = \epsilon^2 I_{M\times M}$ and define $\eta = \epsilon^2/\sigma^2$. Throughout our experiments we take $h = 0.1$, $\epsilon^2 = 0.01$ and fix the parameter $\eta = 0.01$ (i.e. $\sigma = 1$). We use the Lorenz '96 model (3.3.1) to define Ψ, with the parameter choices $F = 8$ and $J = 60$. The system then has 19 positive Lyapunov exponents which we calculate by the methods described in [5]. The observational noise is i.i.d Gaussian with respect to time index k, with distribution $\nu_1 \sim N(0, \epsilon^2)$.

Throughout the following we show (approximation) to the expected value, with respect to noise realizations around a single fixed true signal solving (3.4.2), of the error between the filter and the signal underlying the data, in the Euclidean norm, as a function of time. We also quote numbers which are found by time-averaging this quantity. The expectation is approximated by a Monte Carlo method in which I realizations of the noise in the data are created, leading to filters

$m_k^{(i)}$, with k denoting time and i denoting realization. Thus we have, for $t_k = kh$,

$$\text{RMSE}(t_k) = \frac{1}{I}\sum_{i=1}^{I}\sqrt{\frac{\|m_k^{(i)} - v_k\|^2}{J}}.$$

This quantity is graphed, as a function of k, in what follows. Notice that similar results are obtained if only one realization is used ($I = 1$) but they are more noisy and hence the trends underlying them are not so clear. We take $I = 10^4$ throughout the reported numerical results. When we state a number for the RMSE this will be found by time-averaging after ignoring the initial transients ($t_k < 40$):

$$\text{RMSE} = \underset{t_k>40}{\text{mean}}\{\text{RMSE}(t_k)\}.$$

In what follows we will simply refer to RMSE ; from the context it will be clear whether we are talking about the function of time, RMSE(t_k), or the time-averaged number RMSE.

Figures 3.5.1 and 3.5.2 exhibit, for fixed observation 3DVAR and adaptive observation 3DVAR, the RMSE as a function of time. The Figure 3.5.1 shows the RMSE for fixed observation operator where the observed space is of dimension 60 (complete observations), 40 (observation operator defined as in the equation (3.4.1)), 36 and 24 respectively. For values $M = 60$, 40 and 36 the error decreases rapidly and the approximate solution converges to a neighbourhood of the true solution where the size of the neighbourhood depends upon the variance of the observational noise. For the cases $M = 60$ and $M = 40$ we use the identity operator $I_{J\times J}$ and the projection operator P as defined in the equation (3.4.1) as the observation operators respectively. The observation operator for the case $M = 36$ can be given as

$$P_{36} = \Big(\ e_1,\, e_2,\, 0,\, e_4,\, 0,\, e_6, e_7,\, 0,\, e_9,\, 0,\, e_{11},\, e_{12},\, 0,\, e_{14}, \dots \ \Big)_{J\times J} \tag{3.5.6}$$

where we observe 3 out of 5 directions periodically. The RMSE , averaged over the trajectory, after ignoring the initial transients, is 1.30×10^{-2} when $M = 60$, 1.14×10^{-2} when $M = 40$ and 1.90×10^{-2} when $M = 36$; note that this is on the scale of the observational noise. The rate of convergence of the approximate solution to the true solution in the case of partial observations is lower than the rate of convergence when full observations are used however the RMSE is lower in the case when $M = 40$ due to fewer noisy inputs in stable directions in comparison to the case when all directions are observed. The convergence of the approximate solution to the true solution for the case when $M = 36$ shows that the value $M = 40$, for which theoretical results have been presented in section 3.4, is not required for small error ($\mathcal{O}(\epsilon)$) consistently over the trajectory. We also consider the case when $24 = 40\%$ of the modes are observed using the following observation operator:

$$P_{24} = \Big(\ e_1,\, 0,\, 0,\, e_4,\, 0,\, 0, e_7,\, 0,\, 0,\, e_{10},\, e_{11},\, 0,\, 0,\, e_{14}, \dots \ \Big)_{J\times J}. \tag{3.5.7}$$

Thus we observe 4 out of 10 directions periodically; this structure is motivated by the work reported in [1, 38] where it was demonstrated that observing 40% of the modes, with the observation directions chosen carefully and with observations sufficiently frequent in time, is sufficient for the approximate solution to converge to the true underlying solution. However, in this case the structure of the observation operator is not apparent and the combination of which 24 modes are to be observed is found by trial and error. The Figure 3.5.1 shows that, in our observational set-up, observing 24 of the modes only allows marginally successful reconstruction of the signal, asymptotically in time; the RMSE makes regular large excursions and the time-averaged RMSE over the trajectory is (5.73×10^{-2}), which is an order of magnitude larger than for 36, 40 or 60 observations.

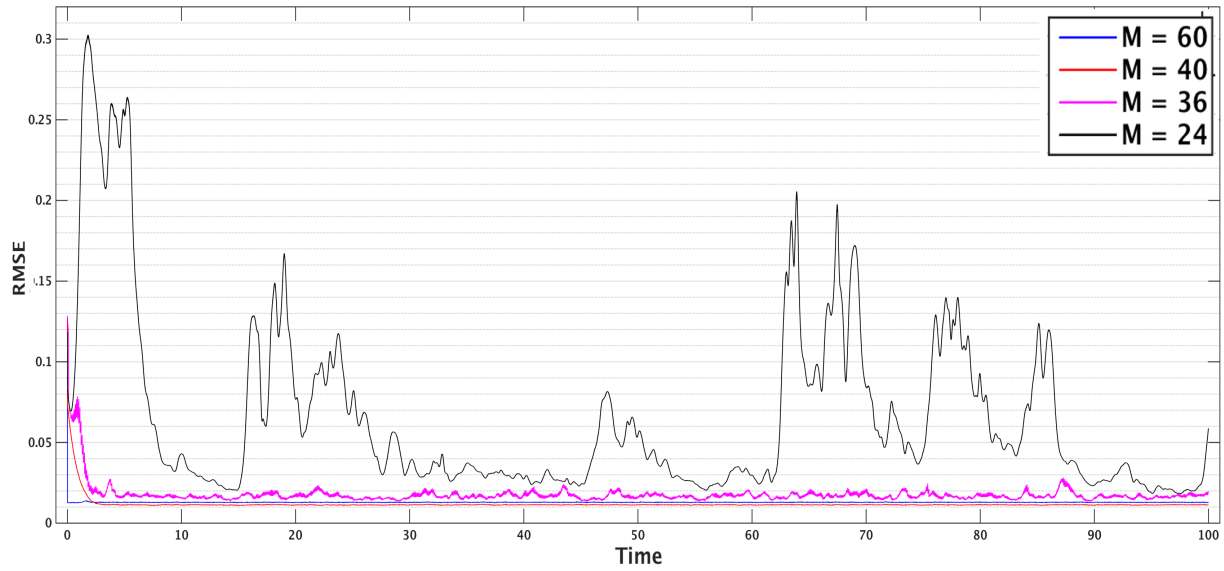
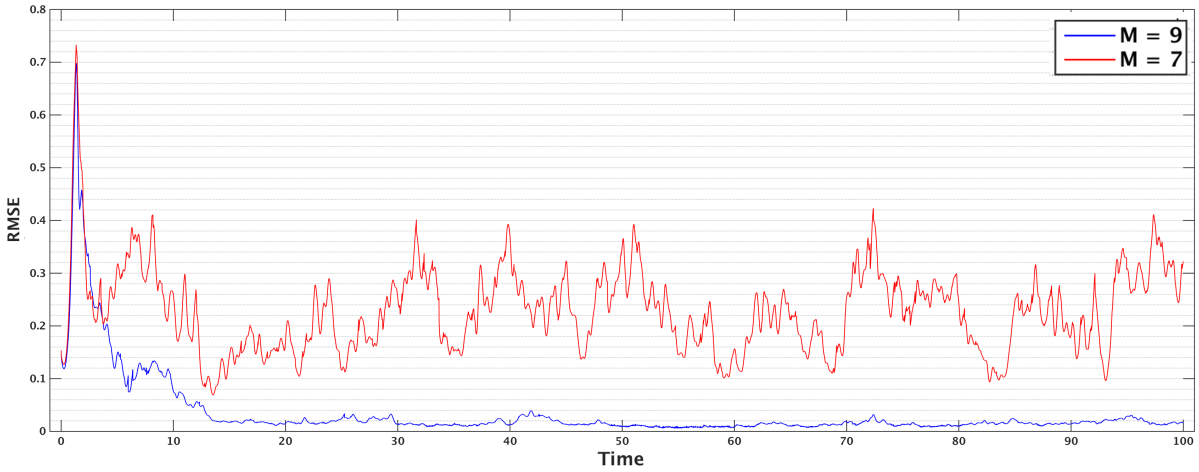


Figure 3.5.1: Fixed Observation Operator 3DVAR. Comparison with the case when $M = 24$. RMSE value averaged over the trajectory for $M = 24$ is 5.73×10^{-2}.

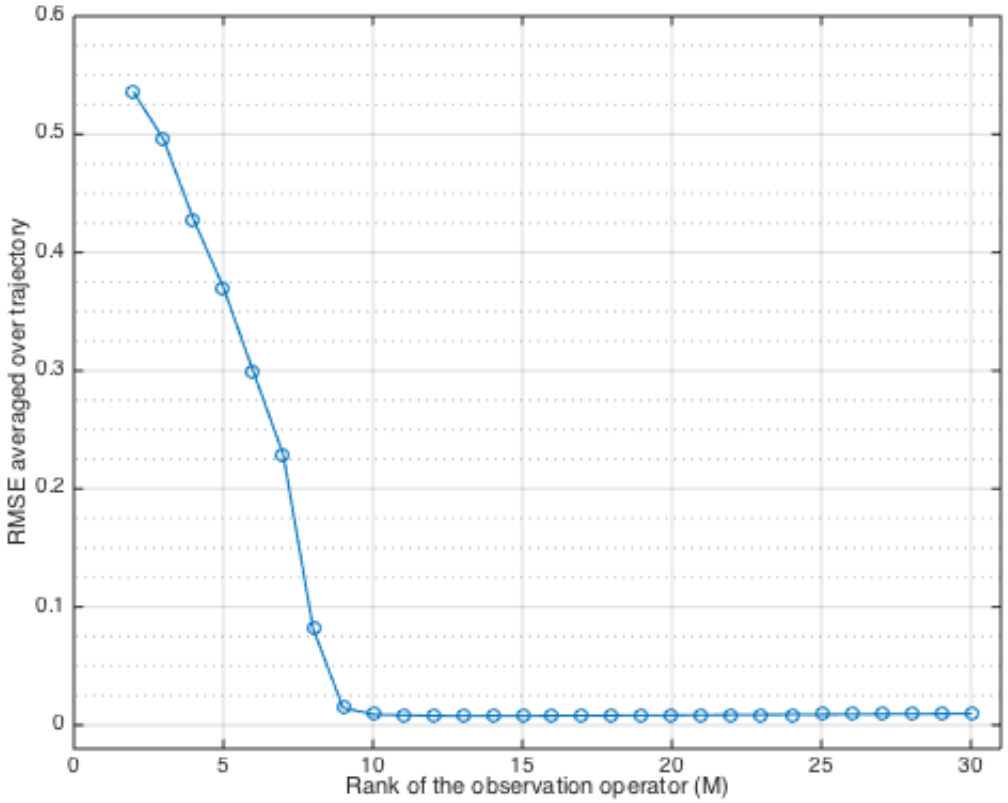
Figure 3.5.2 shows the RMSE for adaptive observation 3DVAR. In this case we notice that the error is consistently small, uniformly in time, with just 9 or more modes observed. When $M = 9$ (15% observed modes) the RMSE averaged over the trajectory is 1.35×10^{-2} which again is of the order of the observational noise variance. For $M \geq 9$ the error is similar – see Figure 3.5.2b. On the other hand, for smaller values of M the error is not controlled as shown in Figure 3.5.2a where the RMSE for $M = 7$ is compared with that for $M = 9$; for $M = 7$ it is an order of magnitude larger than for $M = 9$. It is noteworthy that the number of observations necessary and sufficient for accurate reconstruction is approximately half the number of positive Lyapunov exponents.

3.5.2 Extended Kalman Filter

In the Extended Kalman Filter (ExKF) the approximate solution evolves according to the minimization principle (3.2.4) with C_k chosen as a covariance matrix evolving in the forecast step



(a) Comparison of RMSE between M = 7 and M = 9. RMSE values averaged over trajectory are 2.25×10^{-1}, 1.35×10^{-2} respectively.



(b) Averaged RMSE for different choices of M

Figure 3.5.2: Adaptive Observation 3DVAR

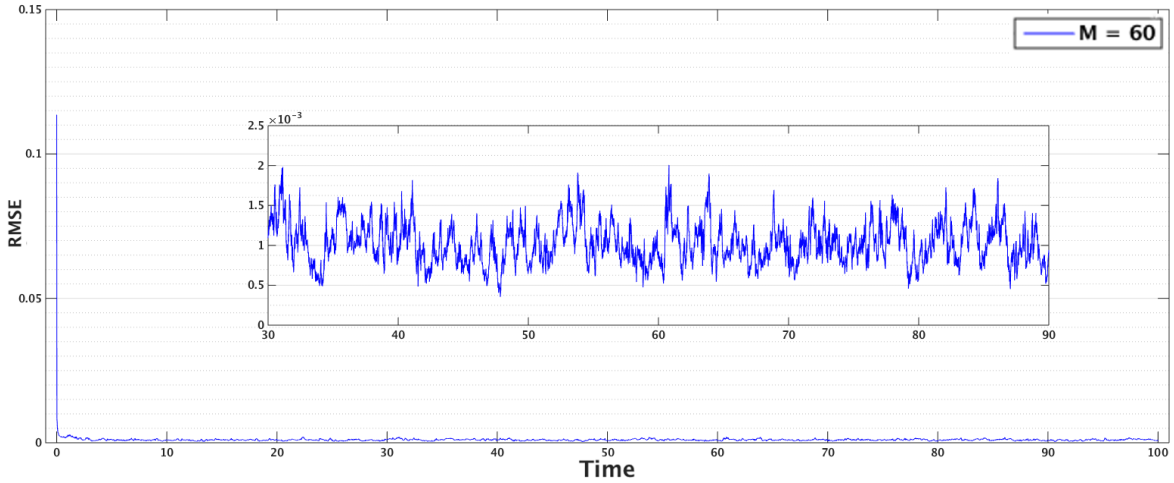
according to the linearized dynamics, and in the assimilation stage updated according to Bayes' rule based on a Gaussian observational error covariance. This gives the method

$$\begin{aligned} m_{k+1} &= \Psi(m_k) + G_{k+1}\Big(y_{k+1} - H_{k+1}\Psi(m_k)\Big), \\ \widehat{C}_{k+1} &= D\Psi(m_k)C_k D\Psi(m_k)^T, \\ C_{k+1} &= (I_{J\times J} - G_{k+1}H_{k+1})\widehat{C}_{k+1}, \\ G_{k+1} &= \widehat{C}_{k+1}H_{k+1}^T(H_{k+1}\widehat{C}_{k+1}H_{k+1}^T + \Gamma)^{-1}. \end{aligned}$$

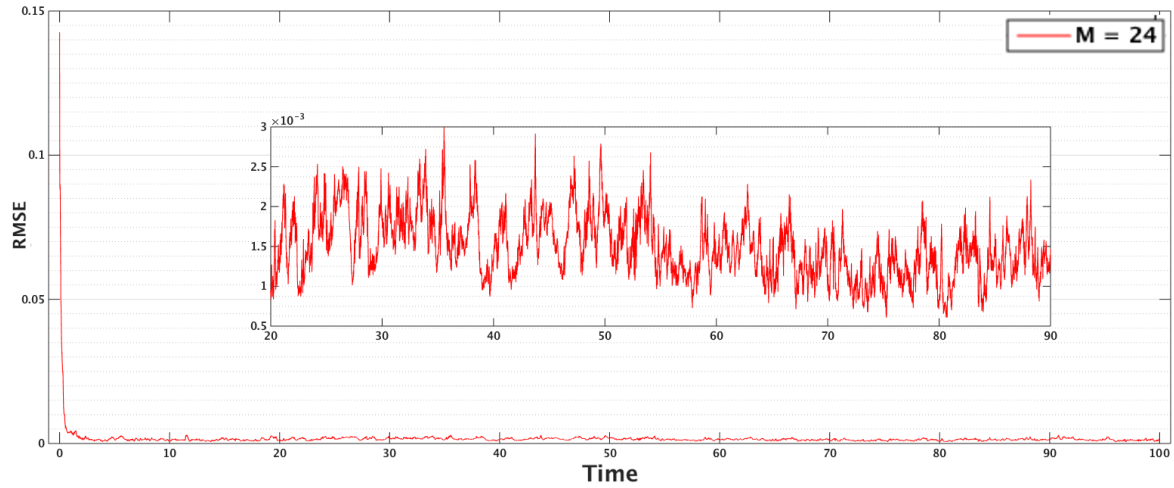
We first consider the ExKF scheme with a fixed observation operator $H_k = H$. We make two choices for H: the full rank identity operator and a partial observation operator given by (3.5.7) so that 40% of the modes are observed. For the first case the filtering scheme is the standard ExKF with all the modes being observed. The approximate solution converges to the true solution and the error decreases rapidly as can be seen in the Figure 3.5.3a. The RMSE is 9.49×10^{-4} which is an order of magnitude smaller than the analogous error for the 3DVAR algorithm when fully observed which is, recall, 1.30×10^{-2}. For the partial observations case with $M = 24$ we see that again the approximate solution converges to the true underlying solution as shown in the Figure 3.5.3b. Furthermore the solution given by the ExKF with $M = 24$ is far more robust than for 3DVAR with this number of observations. The RMSE is also lower for ExKF (2.68×10^{-3}) when compared with the 3DVAR scheme (5.73×10^{-2}).

We now turn to adaptive observation within the context of the ExKF. The Figure 3.5.4 shows that it is possible to obtain an RMSE which is of the order of the observational error, and is robust over long time intervals, using only a 7 dimensional observation space, improving marginally on the 3DVAR situation where 9 dimensions were required to attain a similar level of accuracy.

The AUS scheme, as proposed by Trevisan and co-workers [78], is an ExKF method which operates by confining the analysis update to the subspace spanned by a finite number of directions, ideally designed to capture the instabilities in the dynamics. This is typically achieved by choosing to work in the subspace of the linear dynamics spanned by the M largest growth directions; furthermore M is fixed as the number (precomputed) of non-negative Lyapunov exponents. Asymptotically this method with $H = I_{J\times J}$ behaves similarly to the adaptive ExKF with observation operator of rank M. To understand the intuition behind the AUS method we plot in Figure 3.5.5a the rank (computed by truncation to zero of eigenvalues below a threshold) of the covariance matrix C_k from standard ExKF based on observing 60 and 24 modes. Notice that in both cases the rank approaches a value of 19 or 20 and that 19 is the number of non-negative Lyapunov exponents. This means that the covariance is effectively zero in 40 of the observed dimensions and that, as a consequence of the minimization principle (3.2.4), data will be ignored in the 40 dimensions where the covariance is negligible. It is hence natural to simply confine the update step to the subspace of dimension 19 given by the number of positive Lyapunov exponents, right from the outset. This is exatcly what AUS does by reducing the rank of the error covariance matrix C_k. Numerical results are given in Figure 3.5.5b which shows the RMSE over the trajectory for the ExKF-AUS

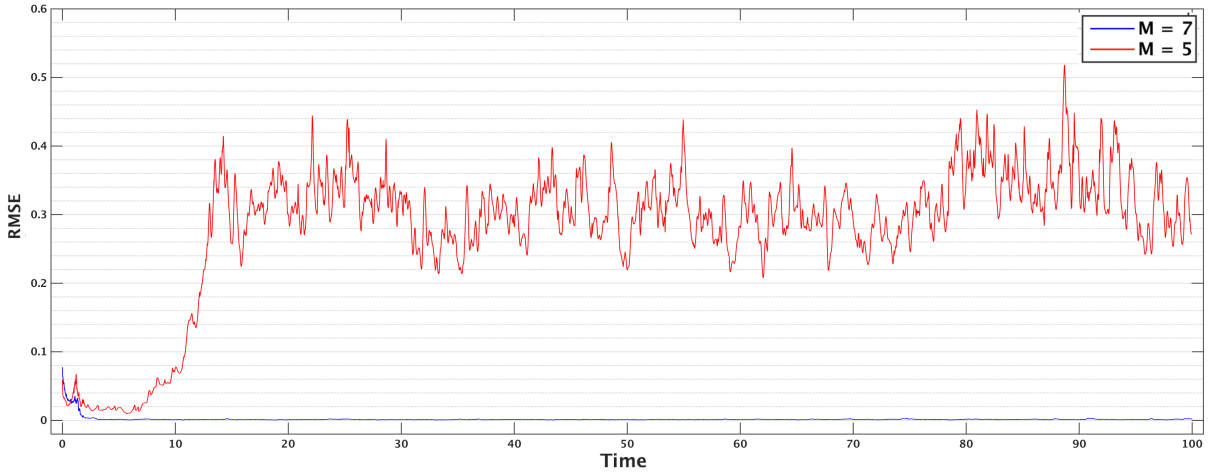


(a) Percentage of components observed = 100%. RMSE value averaged over trajectory 9.49×10^{-4}.

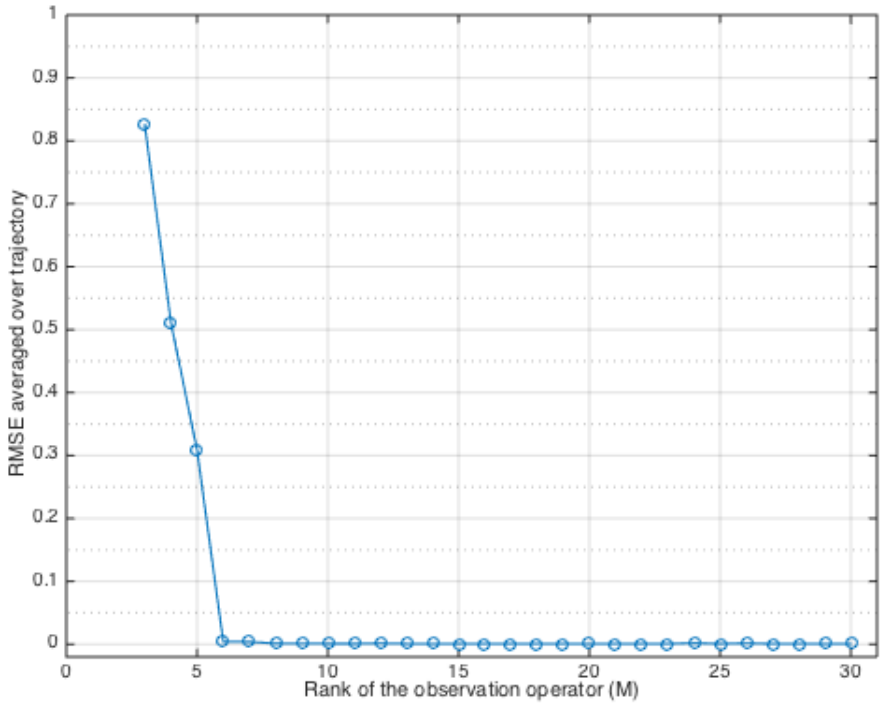


(b) Percentage of components observed = 40%. RMSE value averaged over trajectory 1.39×10^{-3}.

Figure 3.5.3: Fixed Observation ExKF. The zoomed in figures shows the variability in RMSE between time $t = 20$ and $t = 90$.



(a) Comparison of RMSE between M = 5 and M = 7. RMSE values averaged over trajectory are 2.84×10^{-1}, 1.31×10^{-3} respectively.



(b) Averaged RMSE for different choices of M.

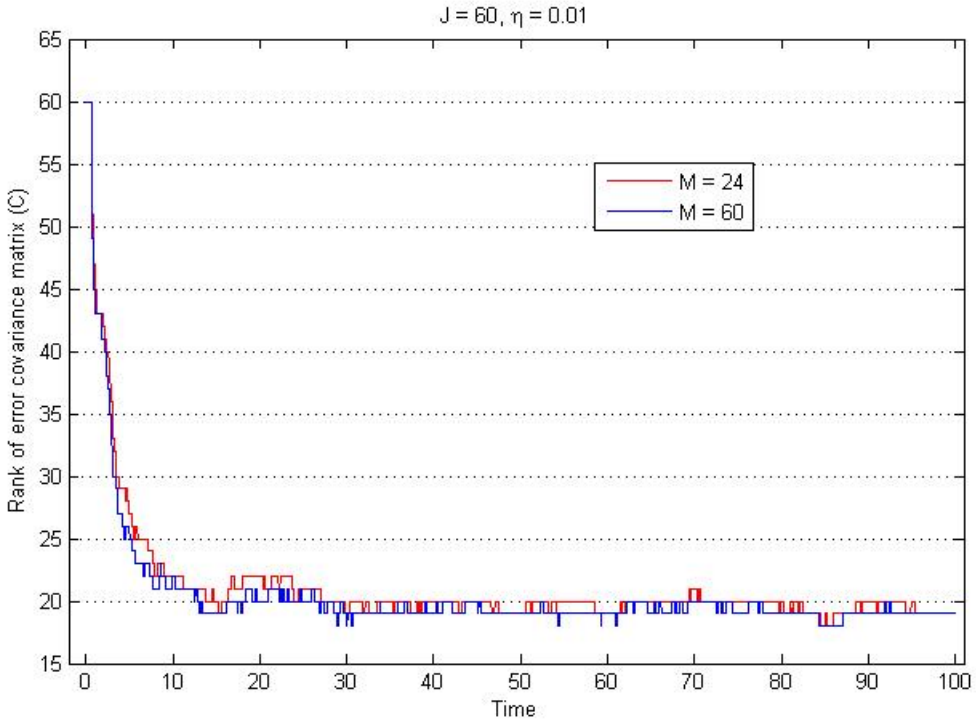
Figure 3.5.4: Adaptive Observation ExKF

assimilation scheme with time for the observation operator $H = I_{J\times J}$. After initial transients the error is mostly of the numerical order of the observational noise. Occasional jumps outside this error bound are observed but the approximate solution converges to the true solution each time. The RMSE for ExKF-AUS is 1.49×10^{-2}. However, if the rank of the error covariance matrix C_0 in AUS is chosen to be less than the number of unstable modes for the underlying system, then the approximate solution does not converge to the true solution.

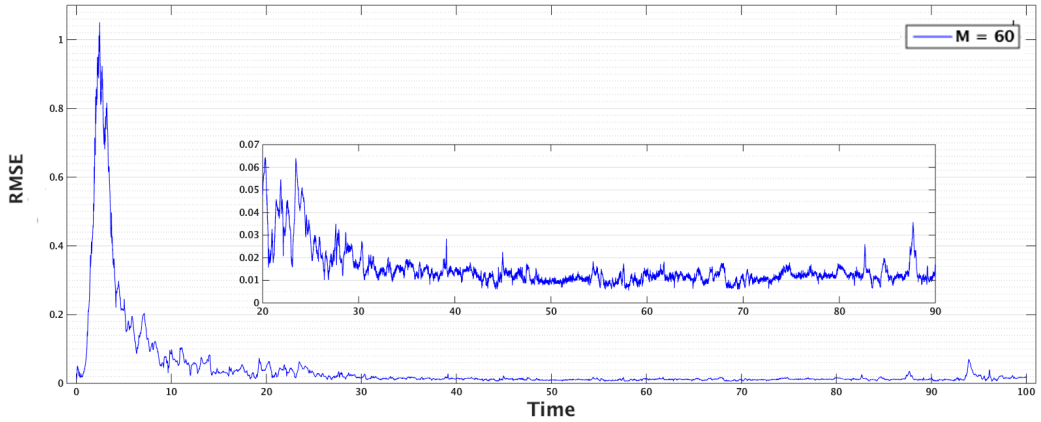
3.6 Conclusions

In this paper we have studied the long-time behaviour of filters for partially observed dissipative dynamical systems, using the Lorenz '96 model as a canonical example. We have highlighted the connection to synchronization in dynamical systems, and shown that this synchronization theory, which applies to noise-free data, is robust to the addition of noise, in both the continuous and discrete time settings. In so doing we are studying the 3DVAR algorithm. In the context of the Lorenz '96 model we have identified a fixed observation operator, based on observing 2/3 of the components of the signal's vector, which is sufficient to ensure desirable long-time properties of the filter. However, it is to be expected that, within the context of fixed observation operators, considerably fewer observations may be needed to ensure such desirable properties. Ideas from nonlinear control theory will be relevant in addressing this issue. We also studied adaptive observation operators, targeted to observe the directions of maximal growth within the local linearized dynamics. We demonstrated that with these adaptive observers, considerably fewer observations are required. We also made a connection between these adaptive observation operators, and the AUS methodology which is also based on the local linearized dynamics, but works by projecting within the model covariance operators of ExKF, whilst the observation operators themselves are fixed; thus the model covariances are adapted. Both adaptive observation operators and the AUS methodology show the potential for considerable computational savings in filtering, without loss of accuracy.

In conclusion, our work highlights the role of ideas from dynamical systems in the rigorous analysis of filtering schemes and, through computational studies shows the gap between theory and practice, demonstrating the need for further theoretical developments. We emphasize that the adaptive observation operator methods may not be implementable in practice on the high dimensional systems arising in, for example, meteorological applications. However, they provide conceptual insights into the development of improved algorithms and it is hence important to understand their properties.



(a) Standard ExKF with 60 and 24 observed modes. The rank of the error covariance matrix C_k decays to (approximately) the number of unstable Lyapunov modes in the underlying system, namely 19.



(b) RMSE value averaged over trajectory: 1.49×10^{-2}. The zoomed in figures shows the variability in RMSE between time $t = 20$ and $t = 90$. The rank of observation operator is chosen $M = 60$.

Figure 3.5.5: Rank of error covariance and ExKF-Assimilation in Unstable Space

Appendix: Proofs

Proof of Properties 3.3.1. Properties 1, 2 and 3 are straightforward and we omit the proofs. We start showing 4. For any $u \in \mathbb{R}^J$ set

$$\|u\|_\infty = \max_{1\le j\le J} |u^{(j)}|$$

and recall that $|u|^2 \ge \|u\|_\infty^2$. Then, for $u, \tilde{u} \in \mathbb{R}^J$, and for $1 \le j \le J$, we have that

$$2|B(u,\tilde{u})^{(j)}| \;\le\; \|u\|_\infty(|\tilde{u}^{(j+1)}| + |\tilde{u}^{(j-2)}|) + \|\tilde{u}\|_\infty(|u^{(j+1)}| + |u^{(j-2)}|),$$

and so

$$\begin{aligned} 4|B(u,\tilde{u})|^2 &\le 2\|u\|_\infty^2 \sum_{j=1}^{J}(|\tilde{u}^{(j+1)}| + |\tilde{u}^{(j-2)}|)^2 + 2\|\tilde{u}\|_\infty^2 \sum_{j=1}^{J}(|u^{(j+1)}| + |u^{(j-2)}|)^2 \\ &\le 8\|u\|_\infty^2|\tilde{u}|^2 + 8\|\tilde{u}\|_\infty^2|u|^2 \\ &\le 16|u|^2|\tilde{u}|^2. \end{aligned}$$

Hence

$$|B(u,\tilde{u})| \le 2|u||\tilde{u}|.$$

For 5 we use rearrangement and periodicity of indices under summation as follows:

$$\begin{aligned} 2\langle B(u,\tilde{u}), u\rangle &= \sum_{j=1}^{J}\Big(u^{(j)}(u^{(j-1)}\tilde{u}^{(j+1)} + \tilde{u}^{(j-1)}u^{(j+1)} - \tilde{u}^{(j-1)}u^{(j-2)} - u^{(j-1)}\tilde{u}^{(j-2)})\Big) \\ &= \sum_{j=1}^{J}(u^{(j)}u^{(j-1)}\tilde{u}^{(j+1)} - u^{(j)}\tilde{u}^{(j-1)}u^{(j-2)}) \\ &= \sum_{j=1}^{J}(u^{(j-1)}u^{(j-2)}\tilde{u}^{(j)} - u^{(j+1)}\tilde{u}^{(j)}u^{(j-1)}) \\ &= \sum_{j=1}^{J}\Big(\tilde{u}^{(j)}(u^{(j-1)}u^{(j-2)} - u^{(j+1)}u^{(j-1)})\Big) \\ &= -\langle B(u,u),\tilde{u}\rangle. \end{aligned}$$

□

Proof of Proposition 3.3.2. Taking the Euclidean inner product of $u(t)$ with equation (3.3.3) and using properties 1 and 2 we get

$$\frac{1}{2}\frac{\mathrm{d}|u|^2}{\mathrm{d}t} = -|u|^2 + \langle f, u\rangle.$$

Using Young's inequality for the last term gives

$$\frac{\mathrm{d}|u|^2}{\mathrm{d}t} + |u|^2 \leq JF^2.$$

Therefore, using Gronwall's lemma,

$$|u(t)|^2 \leq |u_0|^2 e^{-t} + JF^2(1 - e^{-t}),$$

and the result follows. □

Proof of Property 3.4.2. The first part is automatic since, if $q := Qu$, then for all j either $q^{(j-1)} = 0$ or $q^{(j-2)} = q^{(j+1)} = 0$. Since $B(Qu, Qu) = 0$ and $B(\cdot, \cdot)$ is a bilinear operator we can write

$$\begin{aligned} B(u,u) &= B(Pu + Qu, Pu + Qu) \\ &= B(Pu, Pu) + 2B(Pu, Qu). \end{aligned}$$

Now using property 4, and the fact that there is $c > 0$ such that $|Pu| + 2|Qu| \leq \frac{c}{2}|u|$,

$$\begin{aligned} |\langle B(u,u), \tilde{u}\rangle| &\leq |B(u,u)||\tilde{u}| \\ &\leq |B(Pu,Pu) + 2B(Pu,Qu)||\tilde{u}| \\ &\leq 2|Pu||\tilde{u}|(|Pu| + 2|Qu|) \\ &\leq c|Pu||\tilde{u}||u|. \end{aligned}$$

□

Proof of Theorem 3.4.3. Define the error in the approximate solution as $\delta = m - v = q - Qv$. Note that $Q\delta = \delta$. The error satisfies the following equation

$$Q\frac{\mathrm{d}\delta}{\mathrm{d}t} + Q\delta + Q\big(B(Pv + q, Pv + q) - B(v, v)\big) = 0.$$

Splitting $v = Pv + Qv$ and noting, from Properties 3.4.2, that $B(Qv, Qv) = 0$ and $B(q, q) = 0$, yields

$$\frac{\mathrm{d}Q\delta}{\mathrm{d}t} + Q\delta + 2QB(Pv, Q\delta) = 0.$$

Taking the inner product with $Q\delta$ gives

$$\frac{1}{2}\frac{\mathrm{d}|Q\delta|^2}{\mathrm{d}t} + |Q\delta|^2 + 2\langle B(Pv, Q\delta), Q\delta\rangle = 0.$$

Note that from the Properties 3.3.1, 3 and 5, and Property 3.4.2, we have

$$\begin{aligned} 2\langle B(u, Q\delta), Q\delta\rangle &= -\langle B(Q\delta, Q\delta), u\rangle \\ &= 0. \end{aligned}$$

Thus since $Q\delta = \delta$ we have

$$\frac{\mathrm{d}|\delta|^2}{\mathrm{d}t} + 2|\delta|^2 = 0,$$

and so

$$|\delta(t)|^2 = |\delta(0)|^2 e^{-2t}.$$

As $t \to \infty$ the error $\delta(t) \to 0$. □

Proof of Theorem 3.4.4. From (3.4.5) and (3.4.6)

$$\frac{\mathrm{d}m}{\mathrm{d}t} = \mathcal{F}(m) + \frac{1}{\eta}\Big(Pv + \epsilon P\frac{\mathrm{d}w}{\mathrm{d}t} - Pm\Big).$$

Thus

$$\frac{\mathrm{d}m}{\mathrm{d}t} = -m - B(m, m) + f + \frac{1}{\eta}P(v - m) + \frac{\epsilon}{\eta}P\frac{\mathrm{d}w}{\mathrm{d}t}.$$

The signal is given by

$$\frac{\mathrm{d}v}{\mathrm{d}t} = -v - B(v, v) + f,$$

and so the error $\delta = m - v$ satisfies

$$\frac{\mathrm{d}\delta}{\mathrm{d}t} = -\delta - 2B(v, \delta) - B(\delta, \delta) - \frac{1}{\eta}P\delta + \frac{\epsilon}{\eta}P\frac{\mathrm{d}w}{\mathrm{d}t}.$$

Lemma 3.6.2 below, Properties 3.3.1 and Itó's formula give

$$\frac{1}{2}\mathrm{d}|\delta|^2 + \Big(1 - \frac{c^2 K\eta}{4}\Big)|\delta|^2\mathrm{d}t \le \frac{\epsilon}{\eta}\langle P\mathrm{d}w, \delta\rangle + \frac{J}{3}\frac{\epsilon^2}{\eta^2}\mathrm{d}t.$$

Integrating and taking expectations

$$\frac{\mathrm{d}\mathbb{E}|\delta|^2}{\mathrm{d}t} \le -\lambda\mathbb{E}|\delta|^2 + \frac{2J\epsilon^2}{3\eta^2}.$$

Use of the Gronwall inequality gives the desired result. □

We now turn to discrete-time data assimilation, where the following lemma plays an important role:

Lemma 3.6.1. *Consider the Lorenz '96 model* (3.3.3) *with* $F > 0$ *and* $J \ge 3$. *Let* v *and* u *be two*

solutions in $[t_k, t_{k+1})$, *with* $v(t_k) \in \mathcal{B}$. *Then there exists a* $\beta \in \mathbb{R}$ *such that*

$$|u(t) - v(t)|^2 \leq |u(t_k) - v(t_k)|^2 e^{\beta(t-t_k)} \quad t \in [t_k, t_{k+1}).$$

Proof. Let $\delta = m - v$. Then δ satisfies

$$\frac{1}{2}\frac{\mathrm{d}|\delta|^2}{\mathrm{d}t} + |\delta|^2 + 2\langle B(v,\delta),\delta\rangle + \langle B(\delta,\delta),\delta\rangle = 0 \tag{3.6.1}$$

so that, by Property 3.3.1, item 2,

$$\frac{1}{2}\frac{\mathrm{d}|\delta|^2}{\mathrm{d}t} + |\delta|^2 - 2|\langle B(v,\delta),\delta\rangle| \leq 0.$$

Using Properties 3.3.1 items 4 and 5 gives $|\langle B(v,\delta),\delta\rangle| \leq K^{\frac{1}{2}}|\delta|^2$, where K is defined in Proposition 3.3.2, so that

$$\frac{1}{2}\frac{\mathrm{d}|\delta|^2}{\mathrm{d}t} \leq (2K^{\frac{1}{2}} - 1)|\delta|^2.$$

Integrating the differential inequality gives

$$|\delta(t)|^2 \leq |\delta(t_k)|^2 e^{\beta(t-t_k)}. \tag{3.6.2}$$

□

Note if $F < \frac{1}{2\sqrt{2J}}$ then $\beta = 2(2K^{\frac{1}{2}} - 1) < 0$ and the subsequent analysis may be significantly simplified. Thus we assume in what follows that $F \geq \frac{1}{2\sqrt{2J}}$ so that $\beta \geq 0$. Lemma 3.6.1 gives an estimate on the growth of the error in the forecast step. Our aim now is to show that this growth can be controlled by observing Pv discretely in time. It will be required that the time h between observations is sufficiently small.

To ease the notation we introduce three functions that will be used in the proofs of Theorems 3.4.2 and 3.4.6. Namely we define, for $t > 0$,

$$A_1(t) := \frac{16K}{\beta}(e^{\beta t} - 1) + \frac{4R_0^2}{2\beta}(e^{2\beta t} - 1), \tag{3.6.3}$$

$$B_1(t) := \frac{16c^2K^2}{\beta}\left[\frac{e^{\beta t} - e^{-t}}{\beta + 1} - (1 - e^{-t})\right] + e^{-t} + \frac{4c^2KR_0^2}{2\beta}\left[\frac{e^{2\beta t} - e^{-t}}{2\beta + 1} - (1 - e^{-t})\right], \tag{3.6.4}$$

and

$$B_2(t) := c^2K\{1 - e^{-t}\}. \tag{3.6.5}$$

Here and in what follows c, β and K are as in Property 3.4.2, Lemma 3.6.1 and Proposition 3.3.2. We will use two different norms in $\mathbb{R}^J$ to prove the theorems that follow. In each case, the constant $R_0 > 0$ above quantifies the size of the initial error, measured in the relevant norm for the result at hand.

Proof of Theorem 3.4.5. Define the error $\delta = m - v$. Subtracting equation (3.4.2) from equation (3.4.9) gives

$$\frac{\mathrm{d}\delta}{\mathrm{d}t} + \delta + 2B(v,\delta) + B(\delta,\delta) = 0, \quad t \in (t_k, t_{k+1}), \tag{3.6.6a}$$

$$\delta(t_k) = Q\delta(t_k^-) \tag{3.6.6b}$$

where $\delta(t_{k+1}^-) := \lim_{t \uparrow t_{k+1}} \delta(t)$ as defined in section 3.4.2. Notice that $B_1(0) = 1$ and $B_1'(0) = -1$, so that there is $h^* > 0$ with the property that $B_1(h) \in (0,1)$ for all $h \in (0, h^*]$. Fix any such assimilation time h and denote $\gamma = B_1(h) \in (0,1)$. Let $R_0 := |\delta_0|$. We show by induction that, for every k, $|\delta_k|^2 \leq \gamma^k R_0^2$. We suppose that it is true for k and we prove it for $k+1$.

Taking the inner product of $P\delta$ with the equation (3.6.6) gives

$$\frac{1}{2}\frac{\mathrm{d}|P\delta|^2}{\mathrm{d}t} + |P\delta|^2 + 2\langle B(v,\delta), P\delta\rangle + \langle B(\delta,\delta), P\delta\rangle = 0$$

so that, by Property 3.3.1, item 4,

$$\frac{1}{2}\frac{\mathrm{d}|P\delta|^2}{\mathrm{d}t} + |P\delta|^2 \leq 4|v||\delta||P\delta| + 2|\delta|^2|P\delta|.$$

By the inductive hypothesis we have $|\delta_k|^2 \leq R_0^2$ since $\gamma \in (0,1)$. Shifting the time origin by setting $\tau := t - t_k$ and using Lemma 3.6.1 gives

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}|P\delta|^2}{\mathrm{d}\tau} + |P\delta|^2 &\leq 4K^{\frac{1}{2}}|\delta||P\delta| + 2|\delta_k|e^{\frac{\beta\tau}{2}}|\delta||P\delta| \\ &\leq 4K^{\frac{1}{2}}|\delta||P\delta| + 2R_0 e^{\frac{\beta\tau}{2}}|\delta||P\delta|. \end{aligned} \tag{3.6.7}$$

Applying Young's inequality to each term on the right-hand side we obtain

$$\frac{\mathrm{d}|P\delta|^2}{\mathrm{d}\tau} \leq 16K|\delta|^2 + 4R_0^2 e^{\beta\tau}|\delta|^2. \tag{3.6.8}$$

Integrating from 0 to s, where $s \in (0,h)$, gives

$$|P\delta(s)|^2 \leq A_1(s)|\delta_k|^2. \tag{3.6.9}$$

Now again consider the equation (3.6.1) using Property 3.3.1 item 5 to obtain

$$\frac{1}{2}\frac{\mathrm{d}|\delta|^2}{\mathrm{d}\tau} + |\delta|^2 - |\langle B(\delta,\delta), v\rangle| \leq 0.$$

Using Property 3.4.2 and Young's inequality yields

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}|\delta|^2}{\mathrm{d}\tau}+|\delta|^2 &\leq c|v||\delta||P\delta| \\ &\leq cK^{\frac{1}{2}}|\delta||P\delta| \\ &\leq \frac{|\delta|^2}{2}+\frac{c^2K}{2}|P\delta|^2. \end{aligned} \tag{3.6.10}$$

Employing the bound (3.6.9) then gives

$$\frac{\mathrm{d}|\delta|^2}{\mathrm{d}\tau}+|\delta|^2 \leq \left(\frac{16c^2K^2}{\beta}(e^{\beta\tau}-1)+\frac{4c^2KR_0^2}{2\beta}(e^{2\beta\tau}-1)\right)|\delta_k|^2.$$

Therefore, upon using Gronwall's lemma,

$$|\delta(s)|^2 \leq B_1(s)|\delta_k|^2.$$

It follows that

$$|\delta_{k+1}|^2 \leq \gamma|\delta_k|^2 \leq \gamma^{k+1}R_0^2,$$

and the induction (and hence the proof) is complete. □

Proof of Theorem 3.4.6. We define the error process $\delta(t)$ as follows:

$$\delta(t)=\begin{cases} \delta_k := m_k - v(t) & \text{if } t=t_k \\ \Psi(m_k, t-t_k)-v(t) & \text{if } t\in(t_k,t_{k+1}). \end{cases} \tag{3.6.11}$$

Observe that δ is discontinuous at times t_k which are multiples of h, since $m_{k+1}\neq\Psi(m_k;h)$. Subtracting (3.4.11) from (3.4.10) we obtain

$$\delta_{k+1}=\delta(t_{k+1})=\left(\frac{\eta}{1+\eta}P+Q\right)\delta(t_{k+1}^-)+\frac{1}{1+\eta}\nu_{k+1}. \tag{3.6.12}$$

Let $A_1(\cdot)$, $B_1(\cdot)$ and $B_2(\cdot)$ be as in (3.6.3, 3.6.4, 3.6.5), and set

$$\begin{aligned} M_1(t) :=& \frac{2\eta}{1+\eta}\sqrt{A_1(t)}+\sqrt{B_1(t)}, \\ M_2(t) :=& \frac{2\eta}{1+\eta}+\sqrt{B_2(t)}. \end{aligned}$$

Since $A_1(0)=0, B_1(0)=1$, $B_2(0)=0$ and

$$\frac{d}{dt}\sqrt{B_1(t)}\Big|_{t=0}=-1/2<0$$

it is possible to find $h, \eta > 0$ small such that

$$M_2(h) < M_1(h) =: \alpha < 1.$$

Let $R_0 = \|\delta_0\|$. We show by induction that for such h and η, and provided that ϵ is small enough so that

$$\alpha R_0 + 2\epsilon < R_0,$$

we have that $\|\delta_k\| \leq R_0$ for all k. Suppose for induction that it is true for k. Then $|\delta_k| \leq \|\delta_k\| \leq R_0$ and we can apply (after shifting time as before) Lemma 3.6.3 below to obtain that

$$|P\delta(t)| \leq \sqrt{A_1(t)|\delta_k|^2 + |P\delta_k|^2} \leq \sqrt{A_1(t)}|\delta_k| + |P\delta_k|$$

and

$$|\delta(t)| \leq \sqrt{B_1(t)|\delta_k|^2 + B_2(t)|P\delta_k|^2} \leq \sqrt{B_1(t)}|\delta_k| + \sqrt{B_2(t)}|P\delta_k|.$$

Therefore,

$$\begin{aligned} |P\delta_{k+1}| + |\delta_{k+1}| &\leq \left(\frac{2\eta}{1+\eta}\sqrt{A_1(h)} + \sqrt{B_1(h)}\right)|\delta_k| + \left(\frac{2\eta}{1+\eta} + \sqrt{B_2(h)}\right)|P\delta_k| + 2\epsilon \\ &= M_1(h)|\delta_k| + M_2(h)|P\delta_k| + 2\epsilon. \end{aligned}$$

Since $M_2(h) < M_1(h) = \alpha$ we deduce that

$$\|\delta_{k+1}\| \leq \alpha\|\delta_k\| + 2\epsilon,$$

which proves (3.4.12). Furthermore, the induction is complete, since

$$\|\delta_{k+1}\| \leq \alpha\|\delta_k\| + 2\epsilon \leq \alpha R_0 + 2\epsilon \leq R_0.$$

□

Lemma 3.6.2. *Let $v \in \mathcal{B}$. Then, for any δ,*

$$\langle \delta + 2B(v,\delta) + B(\delta,\delta) + \frac{1}{\eta}P\delta, \delta\rangle \geq \left(1 - \frac{c^2 K\eta}{4}\right)|\delta|^2.$$

Proof. Use of Property 3.3.1, items 3 and 5, together with Property 3.4.2, shows that

$$\begin{aligned}\langle \delta + 2B(v,\delta) + B(\delta,\delta) + \frac{1}{\eta}P\delta, \delta\rangle &= |\delta|^2 + 2\langle B(v,\delta),\delta\rangle + \langle B(\delta,\delta),\delta\rangle + \langle \frac{1}{\eta}P\delta,\delta\rangle \\ &= |\delta|^2 - \langle B(\delta,\delta), v\rangle + \langle \frac{1}{\eta}P\delta,\delta\rangle \\ &\geq |\delta|^2 - cK^{\frac{1}{2}}|\delta||P\delta| + \frac{1}{\eta}|P\delta|^2 \\ &\geq |\delta|^2 - \frac{\theta|\delta|^2}{2} - \frac{c^2K|P\delta|^2}{2\theta} + \frac{1}{\eta}|P\delta|^2.\end{aligned}$$

Now choosing $\theta = \frac{c^2K\eta}{2}$ establishes the claim.

□

Lemma 3.6.3. *In the setting of Theorem 3.4.6, for $t \in [0,h)$ and $R_0 := \|\delta_0\|$ we have*

$$|P\delta(t)|^2 \leq A_1(t)|\delta_0|^2 + |P\delta_0|^2 \tag{3.6.13}$$

and

$$|\delta(t)|^2 \leq B_1(t)|\delta_0|^2 + B_2(t)|P\delta_0|^2, \tag{3.6.14}$$

where the error δ is defined as in (3.6.11) *and A_1, B_1 and B_2 are given by (3.6.3, 3.6.4, 3.6.5).*

Proof. As in equation (3.6.8) we have

$$\frac{\mathrm{d}|P\delta|^2}{\mathrm{d}t} \leq 16K|\delta|^2 + 4R_0^2 e^{\beta t}|\delta|^2.$$

On integrating from 0 to t as before, and noting that now $P\delta_0 \neq 0$ in general, we obtain

$$|P\delta(t)|^2 \leq \left(\frac{16K}{\beta}\{e^{\beta t} - 1\} + \frac{4R_0^2}{2\beta}\{e^{2\beta t} - 1\}\right)|\delta_0|^2 + |P\delta_0|^2,$$

which proves (3.6.13).

For the second inequality recall the bound (3.6.10)

$$\frac{1}{2}\frac{\mathrm{d}|\delta|^2}{\mathrm{d}t} + |\delta|^2 \leq \frac{|\delta|^2}{2} + \frac{c^2K}{2}|P\delta|^2,$$

and combine it with (3.6.13) to get

$$\frac{\mathrm{d}|\delta|^2}{\mathrm{d}t} + |\delta|^2 \leq \left(\frac{16c^2K^2}{\beta}\{e^{\beta t} - 1\} + \frac{4c^2KR_0^2}{2\beta}\{e^{2\beta t} - 1\}\right)|\delta_0|^2 + c^2K|P\delta_0|^2.$$

Applying Gronwall's inequality yields (3.6.14). □

Chapter 4

3DVAR constraint 4DVAR Scheme

4.1 Introduction

In the previous chapters we discussed the sequential filtering assimilation schemes. This chapter focuses on the variational assimilation scheme 4DVAR more specifically 3DVAR constraint 4DVAR. Similar to previously discussed 3DVAR, 4DVAR is also a quadratic cost function minimization scheme. In 4DVAR however, a time-sequence of observations can be used by utilising the model dynamics [45, 50, 10, 27]. As with 3DVAR, 4DVAR looks for the analysis as a solution of a minimisation problem, but now using the cost function subject to the strong constraint that the model states must also be a solution of the model equations under consideration. In this chapter we look at the comparative advantages in minimization of 4DVAR cost functions constrained by the Kalman type filters rather than the original model.

The aim of this approach is to use the 3DVAR filter as a constraint to increase the efficacy of the minimization step in 4DVAR scheme. This idea originates from the work on combined filters in [1] and sequential smoothening of posterior variational form in [27]. The aim is to improve the efficacy of numerical minimization for the 4DVAR scheme estimating initial condition. In general if the underlying dynamical system is nonlinear chaotic, the 4DVAR variational form plotted against all possible initial conditions contains multiple local minima which in turn increase the computational cost of the minimization. When using 3DVAR filter as the constraint we observe that the minimization surface smoothens out, although it introduces a bias in the estimate of the initial condition. In this chapter we explore this phenomena both theoretically and numerically.

Throughout this chapter we work with the strong constraint 4DVAR problem for the deterministic systems with perfect model available. For the deterministic systems the aim is to estimate the initial condition first using the model as the constraint and then using 3DVAR filter as the constraint. For the ease of analysis we have considered a linear model as the underlying system, however, the numerical results for both linear and non-linear cases have been presented.

In Section 4.2 we describe the strong constraint 4DVAR formulation and and 3DVAR constraint 4DVAR with respective underlying statistical assumptions. In Section 4.3 we state the

linear dynamical system under consideration and present the results on the presence of bias in the estimation of initial conditions. Section 4.4 takes the path integral approach, discussed in [1], to the bias in the estimation of initial conditions for the underlying system. Numerical experiments are conducted for the linear system in Section 4.5.1 in order to investigate how the theoretical results apply. Further numerical results for Nonlinear systems are presented in Section 4.5.2. Finally we summarize and discuss the principal results of this chapter in Section 4.6.

The main objective of this work is to demonstrate the behaviour of the hybrid estimation scheme under the effect of different parameter regime choices. However in these problems multiple parameters are intricately linked and require in depth study to clearly classify the effects of different parameters on the behaviour of the algorithm which is not the focus of this study. In this work we have stuck to expressing the key behaviours in terms of asymptotic parameter limits and refrained from making assertions about specific effect or combination of effects is responsible for the observed behaviour unless it was apparent from the choice of the parameters (as presented in the case of linear system).

4.2 Set up

Unlike the 3DVAR, where the analysis step only involves the current observation explicitly, 4DVAR is presented as a temporal extension of the 3DVAR for observations that are distributed in time. The basic concept is the same as 3DVAR provided that a nonlinear forecast model is used as part of the observation operator and time integrations of the model are used to provide the model state at the time of the observations. Let v be the solution of the following forward map discretized at times $t_j = jh$ where $0 < j \leq J$ and $h > 0$

$$v_{j+1} = \Psi(v_j) \tag{4.2.1}$$

with the underlying initial condition v_0 and $v_j := v(t_j)$. The discretized map $\Psi(v_j) := \Psi(v_0, jh)$ is a forward integration of the model from time t_0 to time t_j. In the standard formulation of 4DVAR [48, 45], the solution sought is the trajectory of the dynamical model that best fits a series of time-distributed noisy observations of the true state v of the underlying system. The observations $\{y_j\}_{j=1}^J$ are given by

$$y_j = H_j v_j + \nu_j = H_j \Psi^j(v_0) + \nu_j \tag{4.2.2}$$

where the observation errors are i.i.d. random variables distributed as $\nu_j \sim N(0, \Gamma)$, for all j and H_j are the observation operator at time t_j. The objective function, $\mathsf{I}(\cdot) : \mathbb{R}^{J+1} \to \mathbb{R}$, for 4DVAR has the form:

$$\mathsf{I}(u) = \frac{1}{2}\sum_{j=1}^{J} ||y_j - H_j u_j||_\Gamma^2 + \frac{1}{2}||u_0 - m_0||_{C_0}^2. \tag{4.2.3}$$

where the prior mean and variance for the initial condition v_0 are m_0 and C_0 respectively. Substituting the dynamical constraint (4.2.1) into the objective function, the control variable (the trajectory v) is entirely defined by the initial conditions v_0. This is consistent with the perfect model assumption used in this context; that is, the analysis state at the initial time can be integrated with the dynamical model to find the optimal analysis trajectory. In the following we consider the minimization of the reduced objective function $\mathsf{I}_0(\cdot) : \mathbb{R} \to \mathbb{R}$

$$\mathsf{I}_0(u_0) = \frac{1}{2}\sum_{j=1}^{J} ||y_j - H_j\Psi(u_0, jh)||_\Gamma^2 + \frac{1}{2}||u_0 - m_0||_{C_0}^2. \tag{4.2.4}$$

We want to study the case when the state variable u_j is the approximate solution, given by approximate Gaussian data assimilation filtering scheme (e.g. 3DVAR, ExKF or EnKF), which in general sequential form is given by the following equation

$$u_j = \Psi(u_{j-1}) + G_j(y_j - H_j\Psi(u_{j-1})) \tag{4.2.5}$$

where G_j is Kalman gain matrix at j'th assimilation step given by the filter in consideration. We look at the minimization of the functional I where the model state u follows the dynamics described by the equation (4.2.5).

4.3 Linear Case

As a first step to the analysis of the above mentioned algorithm we consider an underlying linear one dimensional system which is modelled by the following equation

$$u_j = \Psi(u_{j-1}) := au_{j-1} \tag{4.3.1}$$

with the initial condition $u_0 \sim N(m_0, C_0 := \sigma^2)$ and $a \in \mathbb{R}^+$. Let v_0 be the initial condition for the true underlying solution v then we can write $v_j = a^j v_0$. The observations of the system are denoted as $\{y_j\}_{j=1}^J$ and defined as

$$y_j = v_j + \nu_j, \tag{4.3.2}$$

where ν_j are i.i.d. random variables distributed as $\nu_j \sim N(0, \Gamma := \epsilon^2)$.

Given the observations we can apply the 3DVAR filtering algorithm which gives us the approximate solution u given by recursively applying the following equation, with the initial condition

u_0,

$$\begin{aligned} u_j &= \Psi(u_{j-1}) + g(y_j - \Psi(u_{j-1})) \\ &= au_{j-1} + g(y_j - au_{j-1}) \\ &= a^j(1-g)^j u_0 + g\sum_{k=1}^{j} a^{j-k}(1-g)^{j-k} y_k \\ &= a^j(1-g)^j u_0 + g\sum_{k=1}^{j} a^{j-k}(1-g)^{j-k}(a^k v_0 + \nu_k) \\ &= a^j(1-g)^j(u_0 - v_0) + a^j v_0 + \sum_{k=1}^{j} a^{j-k}(1-g)^{j-k} g\nu_k \end{aligned} \tag{4.3.3}$$

where g is Kalman gain factor defined as $g := \frac{\sigma^2}{\sigma^2+\epsilon^2}$.

Our aim is to first apply standard strong constraint 4DVAR which correspond to minimizing the cost function I_0 as given by the equation (4.2.4) subject to the hard constraint (4.3.1) and then to consider a modified strong constraint 4DVAR functional by minimizing I as in the equation (4.2.3) subject to the hard constraint (4.3.3).

4.3.1 Problem 1: Standard 4DVAR

The strong 4DVAR minimization problem (4.2.4) with variational form $\mathsf{I}_0(\cdot) : \mathbb{R}^J \to \mathbb{R}$ with constraint equation (4.3.1) for a given set of observations polluted by the noise vector $\{\nu_j\}_{j=1}^J$, where ν_j are i.i.d. random variables distributed as $\nu_j \sim N(0, \epsilon^2)$. Since the forward map in consideration (4.3.1) is deterministic so the minimization over the initial condition is sufficient for the purpose of determining the approximate trajectory. We can write the standard strong constraint 4DVAR cost function I_0 with parameters σ^2 and ϵ^2 as

$$\begin{aligned} \mathsf{I}_0(u_0) &= \frac{1}{2\epsilon^2}\sum_{j=1}^{J} ||y_j - a^j u_0||^2 + \frac{1}{2\sigma^2}||m_0 - u_0||^2 \\ &= \frac{1}{2\epsilon^2}\sum_{j=1}^{J} ||a^j v_0 + \nu_j - a^j u_0||^2 + \frac{1}{2\sigma^2}||m_0 - u_0||^2 \\ &= \frac{1}{2\epsilon^2}\sum_{j=1}^{J} \big(a^j(v_0 - u_0) + \nu_j\big)^2 + \frac{1}{2\sigma^2}(m_0 - u_0)^2 \\ &= \frac{1}{2\epsilon^2}\sum_{j=1}^{J} \big(a^{2j}(v_0 - u_0)^2 + \nu_j^2 + 2a^j(v_0 - u_0)\nu_j\big) + \frac{1}{2\sigma^2}(m_0 - u_0)^2. \end{aligned} \tag{4.3.4}$$

To find the minimizing initial condition we differentiate I_0 w.r.t. u_0

$$\begin{aligned}\frac{\mathrm{d}\mathsf{I}_0(u_0)}{\mathrm{d}u_0} &= -\frac{1}{\epsilon^2}\frac{a^2(a^{2J}-1)}{(a^2-1)}(v_0-u_0) - \frac{1}{\epsilon^2}\sum_{j=1}^{J}a^j\nu_j - \frac{1}{\sigma^2}(m_0-u_0) \\ &= -\frac{1}{\epsilon^2}\frac{a^2(a^{2J}-1)}{(a^2-1)}(v_0-u_0) - \frac{1}{\epsilon^2}\sum_{j=1}^{J}a^j\nu_j - \frac{1}{\sigma^2}(v_0-u_0) - \frac{1}{\sigma^2}(m_0-v_0), \quad (4.3.5)\end{aligned}$$

and set it to 0 and evaluate for the error $(u_0 - v_0)$ as,

$$(u_0-v_0) = \frac{1}{\left(\frac{1}{\sigma^2}+\frac{a^2(a^{2J}-1)}{\epsilon^2(a^2-1)}\right)}\left(\frac{1}{\epsilon^2}\sum_{j=1}^{J}a^j\nu_j + \frac{(m_0-v_0)}{\sigma^2}\right). \quad (4.3.6)$$

With slight misuse of the notation for convenience henceforth we consider u_0 to be the initial condition where minimum of the 4DVAR cost function is observed. We can write down the expression for the squared error in the approximated trajectory as

$$\begin{aligned}(u_0-v_0)^2 &= \frac{1}{\left(\frac{1}{\sigma^2}+\frac{a^2(a^{2J}-1)}{\epsilon^2(a^2-1)}\right)^2}\left(\frac{1}{\epsilon^2}\sum_{j=1}^{J}a^j\nu_j + \frac{(m_0-v_0)}{\sigma^2}\right)^2 \\ (u_0-v_0)^2 &= \frac{1}{\left(\frac{1}{\sigma^2}+\frac{a^2(a^{2J}-1)}{\epsilon^2(a^2-1)}\right)^2}\left(\frac{(m_0-v_0)^2}{\sigma^4} + \frac{1}{\epsilon^4}(\sum_{j=1}^{J}a^j\nu_j)^2 + 2\frac{(m_0-v_0)}{\sigma^2\epsilon^2}\sum_{j=1}^{J}a^j\nu_j\right) \quad (4.3.7)\end{aligned}$$

The above expression is obtained for fixed observational noise vector $\{\nu_j\}_{j=1}^{J}$. To get the mean square error estimate from this expression we could take the expectation considering $\nu_j \sim N(0,\epsilon^2)$. The following theorem presents the main result of this subsection.

Theorem 4.3.1. *Let v be a solution of the linear system given by the equation (4.3.1) with initial condition v_0 and u_0 be the estimate achieved by minimizing the standard strong constraint 4DVAR cost function I_0 with observations as defined in the equation (4.3.2). We observe that if $|a| \geq 1$ then*

$$\lim_{J\to\infty}\mathbb{E}[(u_0-v_0)^2] = 0$$

and if $|a| < 1$ then

$$\lim_{J\to\infty}\mathbb{E}[(u_0-v_0)^2] = \frac{(1-a^2)}{\left((\epsilon^2-\sigma^2)a^2-\epsilon^2\right)^2}\left((m_0-v_0)^2\epsilon^4(1-a^2)+\epsilon^2\sigma^4a^2\right)$$

Proof. We first consider the case when $|a| \neq 1$. Since ν_j are i.i.d. variables with mean 0 and variance ϵ^2 so the noise vector $\{\nu_j\}_{j=1}^{J}$ by taking expectation and using $\mathbb{E}[\nu_j] = 0$ and $\mathbb{E}[{\nu_j}^2] = \epsilon^2$

$$\begin{aligned}
\mathbb{E}[(u_0 - v_0)^2] &= \frac{1}{\left(\frac{1}{\sigma^2} + \frac{a^2(a^{2J}-1)}{\epsilon^2(a^2-1)}\right)^2}\left(\frac{(m_0 - v_0)^2}{\sigma^4} + \frac{1}{\epsilon^4}\mathbb{E}[(\sum_{j=1}^{J} a^j \nu_j)^2] - 2\frac{(m_0 - v_0)}{\sigma^2\epsilon^2}\mathbb{E}[\sum_{j=1}^{J} a^j \nu_j]\right) \\
&= \frac{1}{\left(\frac{1}{\sigma^2} + \frac{a^2(a^{2J}-1)}{\epsilon^2(a^2-1)}\right)^2}\left(\frac{(m_0 - v_0)^2}{\sigma^4} + \frac{\epsilon^2}{\epsilon^4}\sum_{j=1}^{J} a^{2j}\right) \\
&= \frac{1}{\left(\frac{1}{\sigma^2} + \frac{a^2(a^{2J}-1)}{\epsilon^2(a^2-1)}\right)^2}\left(\frac{(m_0 - v_0)^2}{\sigma^4} + \frac{\epsilon^2}{\epsilon^4}\frac{a^2(a^{2J}-1)}{(a^2-1)}\right). \qquad (4.3.8)
\end{aligned}$$

On rearranging terms we get

$$\mathbb{E}[(u_0 - v_0)^2] = \frac{\epsilon^2(a^2-1)}{\left(\sigma^2 a^{2(J+1)} + (\epsilon^2 - \sigma^2)a^2 - \epsilon^2\right)^2}\left((m_0 - v_0)^2\epsilon^2(a^2-1) + \sigma^4 a^2(a^{2J}-1)\right).$$

and taking the limit as $J \to \infty$ gives the required result.

For the case when $|a| = 1$ we get

$$\mathbb{E}[(u_0 - v_0)^2] = \frac{1}{\left(\frac{1}{\sigma^2} + \frac{J}{\epsilon^2}\right)^2}\left(\frac{(m_0 - v_0)^2}{\sigma^4} + \frac{J}{\epsilon^2}\right) \qquad (4.3.9)$$

which also satisfies $\lim_{J\to\infty} \mathbb{E}[(u_0 - v_0)^2] = 0$. □

Remark 4.3.2. *The result indicates that for linear systems if the growth coefficient $|a| \geq 1$, then the 4DVAR filtering system incorporates enough information from the difference between the observations and the background model states (innovation vector) to get the mean square convergence to the underlying true state asymptotically.*

- *In the case of $|a| < 1$ the system decays and the observations are overpolluted by the noise which introduces a non-zero error in the 4DVAR estimate.*
- *The error depends on the size of the observational noise and can be made small in the small noise regime. We notice that as $\epsilon^2 \to 0$ the limit $\lim_{J\to\infty} \mathbb{E}[(u_0 - v_0)^2] \to 0$ even for the case when $|a| < 1$.*
- *The regime when $0 < \sigma^2 \ll 1$, indicates concentration of probability around the prior mean we notice that $\lim_{J\to\infty} \mathbb{E}[(u_0 - v_0)^2] \to (m_0 - v_0)^2$ in the case when $|a| < 1$ as expected.*

4.3.2 Problem 2: 3DVAR constraint 4DVAR

Now we consider the modified strong constraint 4DVAR found by minimizing the variational form I subject to the 3DVAR solution given by (4.3.2) and the observation sequence $\{y_j\}_{j=1}^J$. We can write the standard strong constraint 4DVAR cost function I minimizing over the initial condition with parameters σ^2 and ϵ^2 as

$$\begin{aligned}
\mathsf{I}(u_0) &= \frac{1}{2\epsilon^2}\sum_{j=1}^{J}||y_j - u_j||^2 + \frac{1}{2\sigma^2}||m_0 - u_0||^2 \\
&= \frac{1}{2\epsilon^2}\sum_{j=1}^{J}||a^j v_0 + \nu_j - u_j||^2 + \frac{1}{2\sigma^2}||m_0 - u_0||^2 \\
&= \frac{1}{2\epsilon^2}\sum_{j=1}^{J}||a^j v_0 + \nu_j + a^j(1-g)^j(v_0 - u_0) - a^j v_0 - \sum_{k=1}^{j} a^{j-k}(1-g)^{j-k} g\nu_k||^2 + \frac{1}{2\sigma^2}||m_0 - u_0||^2 \\
&= \frac{1}{2\epsilon^2}\sum_{j=1}^{J}||\nu_j + a^j(1-g)^j(v_0 - u_0) - \sum_{k=1}^{j} a^{j-k}(1-g)^{j-k} g\nu_k||^2 + \frac{1}{2\sigma^2}||m_0 - u_0||^2. \qquad (4.3.10)
\end{aligned}$$

We define $\tilde{a} := a(1-g)$. On differentiating the cost function I_0,

$$\begin{aligned}
\frac{\mathrm{d}\mathsf{I}(u_0)}{\mathrm{d}u_0} &= \Big(-\frac{1}{\epsilon^2}\frac{\tilde{a}^2(\tilde{a}^{2J}-1)}{(\tilde{a}^2-1)}(v_0 - u_0) - \frac{1}{\epsilon^2}\sum_{j=1}^{J}\tilde{a}^j\big(\nu_j - \sum_{k=1}^{j}\tilde{a}^{j-k} g\nu_k\big)\Big) - \frac{1}{\sigma^2}(m_0 - u_0) \\
&= \Big(-\frac{1}{\epsilon^2}\frac{\tilde{a}^2(\tilde{a}^{2J}-1)}{(\tilde{a}^2-1)}(v_0 - u_0) - \frac{1}{\epsilon^2}\sum_{j=1}^{J}\tilde{a}^j\big(\nu_j - \sum_{k=1}^{j}\tilde{a}^{j-k} g\nu_k\big)\Big) - \frac{1}{\sigma^2}(v_0 - u_0) - \frac{1}{\sigma^2}(m_0 - v_0)
\end{aligned}$$

setting it to zero

$$(u_0 - v_0) = \frac{1}{\Big(\frac{1}{\sigma^2} + \frac{1}{\epsilon^2}\frac{\tilde{a}^2(\tilde{a}^{2J}-1)}{(\tilde{a}^2-1)}\Big)}\Big(\frac{1}{\epsilon^2}\sum_{j=1}^{J}\tilde{a}^j\big(\nu_j - \sum_{k=1}^{j}\tilde{a}^{j-k} g\nu_k\big) - \frac{(m_0 - v_0)}{\sigma^2}\Big) \qquad (4.3.11)$$

To calculate the mean square error

$$\begin{aligned}
(u_0 - v_0)^2 &= \frac{1}{\Big(\frac{1}{\sigma^2} + \frac{1}{\epsilon^2}\frac{\tilde{a}^2(\tilde{a}^{2J}-1)}{(\tilde{a}^2-1)}\Big)^2}\Big(\frac{1}{\epsilon^2}\sum_{j=1}^{J}\tilde{a}^j\big((1-g)\nu_j - \sum_{k=1}^{j-1}\tilde{a}^{j-k} g\nu_k\big) - \frac{(m_0 - v_0)}{\sigma^2}\Big)^2 \\
&= \frac{1}{\Big(\frac{1}{\sigma^2} + \frac{1}{\epsilon^2}\frac{\tilde{a}^2(\tilde{a}^{2J}-1)}{(\tilde{a}^2-1)}\Big)^2}\Big(\frac{(m_0 - v_0)^2}{\sigma^4} + \Big(\frac{1}{\epsilon^2}\sum_{j=1}^{J}\tilde{a}^j\big((1-g)\nu_j - \sum_{k=1}^{j-1}\tilde{a}^{j-k} g\nu_k\big)\Big)^2 \\
&- 2\frac{(m_0 - v_0)}{\sigma^2}\Big(\frac{1}{\epsilon^2}\sum_{j=1}^{J}\tilde{a}^j\big((1-g)\nu_j - \sum_{k=1}^{j-1}\tilde{a}^{j-k} g\nu_k\big)\Big)\Big) \qquad (4.3.12)
\end{aligned}$$

Since the above expression is obtained for fixed observational noise vector $\{\nu_j\}_{j=1}^{J}$ we can get the mean square error estimate from this expression by considering $\nu_j \sim N(0, \epsilon^2)$. The corresponding result can be expressed as the following theorem.

Theorem 4.3.3. *Let v be a solution of the linear system given by the equation (4.3.1) with initial condition v_0 and u_0 be the estimate achieved by minimizing the standard strong constraint 4DVAR cost function I with observations as defined in the equation (4.3.2). We observe that if $|\tilde{a}| > 1$ then*

$$\lim_{J\to\infty} \mathbb{E}[(u_0 - v_0)^2] = \frac{\epsilon^2\sigma^4}{(\tilde{a}^2-1)(\sigma^2+\epsilon^2)^2}$$

and if $|\tilde{a}| < 1$ then

$$\mathbb{E}[(u_0 - v_0)^2] = \frac{1}{\left(\frac{1}{\sigma^2} + \frac{1}{\epsilon^2}\frac{\tilde{a}^2}{(1-\tilde{a}^2)}\right)^2}\left(\frac{(m_0-v_0)^2}{\sigma^4} + \frac{\tilde{a}^2}{\epsilon^2(1-\tilde{a}^2)}\left[(1-g)^2 + \frac{g^2\tilde{a}^4}{(1-\tilde{a}^2)^2} - \frac{2g(1-g)\tilde{a}^2}{(1-\tilde{a}^2)}\right]\right).$$

Proof. On taking the expectation for the squared error

$$\mathbb{E}[(u_0 - v_0)^2] = \frac{1}{\left(\frac{1}{\sigma^2} + \frac{1}{\epsilon^2}\frac{\tilde{a}^2(\tilde{a}^{2J}-1)}{(\tilde{a}^2-1)}\right)^2}\left(\frac{(m_0-v_0)^2}{\sigma^4} + \mathbb{E}\left[\left(\frac{1}{\epsilon^2}\sum_{j=1}^{J}\tilde{a}^j\big((1-g)\nu_j - \sum_{k=1}^{j-1}\tilde{a}^{j-k}g\nu_k\big)\right)^2\right] - 2\frac{(m_0-v_0)}{\sigma^2\epsilon^2}\mathbb{E}\left[\sum_{j=1}^{J}\tilde{a}^j\big((1-g)\nu_j - \sum_{k=1}^{j-1}\tilde{a}^{j-k}g\nu_k\big)\right]\right)$$

since ν_j are i.i.d. variables with mean 0 and variance ϵ^2 so we get

$$\begin{aligned}\mathbb{E}[(u_0 - v_0)^2] &= \frac{1}{\left(\frac{1}{\sigma^2} + \frac{1}{\epsilon^2}\frac{\tilde{a}^2(\tilde{a}^{2J}-1)}{(\tilde{a}^2-1)}\right)^2}\left(\frac{(m_0-v_0)^2}{\sigma^4} + \frac{1}{\epsilon^2}\sum_{j=1}^{J}\left[(1-g)^2\tilde{a}^{2j}\right.\right.\\ &+ \left.\left.\frac{g^2\tilde{a}^4}{(\tilde{a}^2-1)^2}(\tilde{a}^{4J-2j} + \tilde{a}^{2j} - \tilde{a}^{2J}) - \frac{2g(1-g)\tilde{a}^2}{(\tilde{a}^2-1)}(\tilde{a}^{2J} - \tilde{a}^{2j})\right]\right). \quad (4.3.13)\end{aligned}$$

Summing up the series gives the following expression

$$\begin{aligned}\mathbb{E}[(u_0 - v_0)^2] &= \frac{1}{\left(\frac{1}{\sigma^2} + \frac{1}{\epsilon^2}\frac{\tilde{a}^2(\tilde{a}^{2J}-1)}{(\tilde{a}^2-1)}\right)^2}\left(\frac{(m_0-v_0)^2}{\sigma^4} + \frac{1}{\epsilon^2}\left[(1-g)^2\frac{\tilde{a}^2(\tilde{a}^{2J}-1)}{(\tilde{a}^2-1)}\right.\right.\\ &+ \frac{g^2\tilde{a}^4}{(\tilde{a}^2-1)^2}\left(\frac{\tilde{a}^{2J}(\tilde{a}^{2J}-1)}{(\tilde{a}^2-1)} + \frac{\tilde{a}^2(\tilde{a}^{2J}-1)}{(\tilde{a}^2-1)} - J\tilde{a}^{2J}\right)\\ &- \left.\left.\frac{2g(1-g)\tilde{a}^2}{(\tilde{a}^2-1)}\left(J\tilde{a}^{2J} - \frac{\tilde{a}^2(\tilde{a}^{2J}-1)}{(\tilde{a}^2-1)}\right)\right]\right)\end{aligned}$$

and taking the limit as $J \to \infty$ gives the desired result. □

Remark 4.3.4. *Notice that instead of the growth coefficient a, the relevant coefficient is* $\tilde{a}$ *which is defined as* $\tilde{a} := a(1-g) = \frac{a\epsilon^2}{\sigma^2+\epsilon^2}$.

- *If* $\tilde{a} < 1$ *we get a non-zero error depending upon the error in the initial condition which is similar to the case in the theorem 4.3.1. However for the case* $\tilde{a} > 1$ *the result is different in the theorem 4.3.3 compared to the one in the theorem 4.3.1. The mean square error does not go to zero eventually but to a finite limit.*

- *In small noise limit when* $\epsilon^2 \to 0$ *we notice that* $\tilde{a} := \frac{a\epsilon^2}{\sigma^2+\epsilon^2} \to 0$ *and the limit* $\lim_{J\to\infty} \mathbb{E}[(u_0 - v_0)^2] \to (m_0 - v_0)^2$ *so unlike the standard 4DVAR, 3DVAR constraint 4DVAR introduces a bias depending upon the prior mean. Similar result is also observed in the case when the prior variance is large as when* $\sigma^2 \to \infty$ *we get* $\tilde{a} \to 0$.

- *In the case when* $0 < \sigma^2 \ll 1$ *the algorithm effectively ignores the 3DVAR innovation term and the constrain on the variational term is similar to the standard 4DVAR algorithm. In this case when* $|\tilde{a}| > 1$ *then* $\lim_{J\to\infty} \mathbb{E}[(u_0 - v_0)^2] \to 0$ *and* $\lim_{J\to\infty} \mathbb{E}[(u_0 - v_0)^2] \to (m_0 - v_0)^2$ *when* $|\tilde{a}| < 1$ *similar to the case in the remark 4.3.2.*

4.4 Calculations with Posterior Distribution

In this section we take the approach of formulating path integrals for the linear dynamical system described in section 4.3 to estimate the true initial condition given discrete noisy observations of the system. We calculate the cumulant generating function, as explained in [65], for the posterior distribution of the initial condition u_0 given the data $\{y_j\}_{j=1}^J$ and corresponding objective function i.e. standard or constraint 4DVAR. This method provides us the mean value of initial condition for the posterior distribution for fixed set of observations and also allows us probabilistic underpinnings when considered over the space of all possible noise realizations and infinitely large number of observations are allowed.

4.4.1 Standard 4DVAR

We first consider the linear system defined by the equation (4.3.1) where we assume the prior distribution for the initial condition u_0 to be Gaussian given as $N(m_0, C_0)$. The discrete noisy observations for the system are provided as in the equation (4.3.2) where additive Gaussian noise is distributed as $N(0, \epsilon^2)$. To calculate the posterior distribution we require the likelihood function of the observed data which can be given as

$$\mathcal{L}(\{y_j\}_{j=1}^J | u_0) = e^{-\frac{1}{2}\sum_{j=1}^{J} ||y_j - u_j||^2_{\epsilon^2}}$$

where $\epsilon^2 > 0$. Now we can write the posterior distribution function as

$$P(u_0|\{y_j\}_{j=1}^J) = \frac{1}{Z} e^{-\frac{1}{2}||u_0 - m_0||^2_{C_0}} e^{-\frac{1}{2}\sum_{j=1}^{J} ||y_j - u_j||^2_{\epsilon^2}} \tag{4.4.1}$$

where Z is normalizing constant and u_j is given by the equation (4.3.1). The cumulant generating function for the above posterior distribution can be written as following

$$e^{C(\kappa)} = \frac{1}{Z} \int e^{\kappa u_0} e^{-\frac{1}{2}||u_0 - m_0||^2_{C_0}} e^{-\frac{1}{2}\sum_{j=1}^{J} ||y_j - a^j u_0||^2_{\epsilon^2}} \mathrm{d}u_0. \tag{4.4.2}$$

On expanding the terms and collecting the powers of u_0 we get

$$\begin{aligned} e^{C(\kappa)} &= \frac{1}{Z} \int e^{\kappa u_0} e^{-\frac{1}{2}||u_0 - m_0||^2_{C_0}} e^{-\frac{1}{2}\sum_{j=1}^{J} ||y_j - a^j u_0||^2_{\epsilon^2}} \mathrm{d}u_0 \\ &= \frac{1}{Z} \int e^{-\alpha u_0^2 + \beta u_0 + c} \mathrm{d}u_0 \end{aligned} \tag{4.4.3}$$

where α, β and c are defined as

$$\alpha = \frac{1}{2C_0} + \frac{1}{2\epsilon^2} \sum_{j=1}^{J} a^{2j} \tag{4.4.4}$$

$$\beta = \kappa + \frac{m_0}{C_0} + \frac{1}{\epsilon^2} \sum_{j=1}^{J} a^j y_j \tag{4.4.5}$$

$$c = -\frac{m_0^2}{2C_0} - \frac{1}{2\epsilon^2} \sum_{j=1}^{J} y_j^2. \tag{4.4.6}$$

Since C is the cumulant generating function for the posterior distribution we get the following relations

$$\mathbb{E}_P[u_0] = \left.\frac{\partial C}{\partial \kappa}\right|_{\kappa=0} \tag{4.4.7}$$

$$\mathbb{E}_P[u_0^2] - \mathbb{E}_P[u_0]^2 = \left.\frac{\partial^2 C}{\partial \kappa^2}\right|_{\kappa=0}. \tag{4.4.8}$$

To calculate these values we take natural Logarithm of both sides in equation (4.4.3)

$$\ln e^{C(\kappa)} = \ln\left(e^{\frac{\beta^2}{4\alpha} + c} \int e^{-(\sqrt{\alpha}u_0 - \frac{\beta}{2\sqrt{\alpha}})^2} \mathrm{d}u_0\right) \tag{4.4.9}$$

$$C(\kappa) = \frac{\beta^2}{4\alpha} + c = \frac{1}{4\alpha}\left(\kappa^2 + 2\kappa\left(\frac{m_0}{C_0} + \frac{1}{\epsilon^2}\sum_{k=1}^{J} a^j y_j\right)\right) \underbrace{-\frac{m_0^2}{2C_0} - \frac{1}{\epsilon^2}\sum_{j=1}^{J} y_j^2}_{\text{terms independent of } \kappa}. \tag{4.4.10}$$

On differentiating the cumulant generating function $C(\kappa)$ we get

$$\left.\frac{\partial C}{\partial \kappa}\right|_{\kappa=0} = \frac{1}{2\alpha}\left(\frac{m_0}{C_0} + \frac{1}{\epsilon^2}\sum_{j=1}^{J} a^j y_j\right) \tag{4.4.11}$$

and differentiating again gives

$$\left.\frac{\partial^2 C}{\partial \kappa^2}\right|_{\kappa=0} = \frac{1}{4\alpha}\frac{\partial^2 \beta^2}{\partial \kappa^2} = \frac{1}{2\alpha}. \tag{4.4.12}$$

Substituting these values in the equations (4.4.7) and (4.4.8) gives

$$\mathbb{E}_P[u_0] = \frac{1}{2\alpha}(\frac{m_0}{C_0} + \frac{1}{\epsilon^2}\sum_{j=1}^{J} a^j y_j), \tag{4.4.13}$$

$$\mathbb{E}_P[u_0^2] - \mathbb{E}_P[u_0]^2 = \frac{1}{2\alpha}. \tag{4.4.14}$$

Before proving the main result of this section we show that $\lim_{J\to\infty}\left(\frac{1}{S^J(a)}\sum_{j=1}^{J} a^j \nu_j\right) = 0$ almost surely, where we define $S^J(a) := \frac{a^2(a^{2J}-1)}{a^2-1}$.

Lemma 4.4.1. *Let ν_j be i.i.d. Gaussian random variables with mean 0 and variance ϵ^2 and $a > 1$ be a constant then*

$$\lim_{J\to\infty}\left(\frac{1}{S^J(a)}\sum_{j=1}^{J} a^j \nu_j\right) = 0$$

almost surely.

Proof. Since the random variables $\nu_j \sim N(0, \epsilon^2)$ for all j, the random variables $a^j\nu_j \sim N(0, a^{2j}\epsilon^2)$ for all j. Now consider the sum $S^J := \sum_{j=1}^{J} a^j \nu_j$ of independent Gaussian random variables, we know that the sum of independent Gaussian random variable is also Gaussian which gives $S^J \sim N\left(0, \epsilon^2\sum_{j=1}^{J} a^{2j}\right)$. Now we use the Chebyshev's inequality as, for given $\delta > 0$

$$\Pr\big[|S^J| \geq \delta S^J(a)\big] \leq \frac{\epsilon^2}{\delta^2 S^J(a)^2}\sum_{j=1}^{J} a^{2j} = \frac{\epsilon^2}{\delta^2 S^J(a)}.$$

Now since $a > 1$ we know $\lim_{J\to\infty} S^J(a) = \infty$ so as $J \to \infty$

$$\Pr\left[\left|\frac{S^J}{S^J(a)}\right| \geq \delta\right] \to 0,$$

This gives us that $\frac{S^J}{S^J(a)} \to 0$ in probability. Now we define $J_j := \inf\{J : S^J(a) \geq j^2\}$ then we have

$$\Pr\left[\left|\frac{S^{J_j}}{S^{J_j}(a)}\right| \geq \delta\right] \leq \frac{\epsilon^2}{\delta^2 S^{J_j}(a)} \leq \frac{\epsilon^2}{\delta^2 j^2}.$$

On summing over j we get

$$\sum_{j=1}^{\infty}\Pr\left[\left|\frac{S^{J_j}}{S^{J_j}(a)}\right| \geq \delta\right] \leq \sum_{j=1}^{\infty}\frac{\epsilon^2}{\delta^2 j^2} \leq \infty.$$

Now by applying Borel-Cantelli lemma we get $\frac{S^{J_j}}{S^{J_j}(a)} \to 0$ almost surely as $j \to \infty$. Since

$$\frac{S^{J_j}}{S^{J_{j+1}}(a)} \leq \frac{S^J}{S^J(a)} \leq \frac{S^{J_{j+1}}}{S^{J_j}(a)}$$

for all $J_j < J \leq J_{j+1}$ and $\frac{S^{J_{j+1}}(a)}{S^{J_j}(a)} \to 1$ as $j \to \infty$ which gives $\frac{S^J}{S^J(a)} \to 0$ almost surely. □

Theorem 4.4.2. *Let v be a solution of the linear system given by the equation* (4.3.1) *with initial condition v_0 and u_0 be the maximum a posteriori estimate for the distribution given by the equation* (4.4.1)*, with observations as defined in the equation* (4.3.2)*. We observe that if $a > 1$ then $\lim_{J\to\infty} \mathbb{E}_P[u_0] \to v_0$ almost surely and $\lim_{J\to\infty} (\mathbb{E}_P[u_0^2] - \mathbb{E}_P[u_0]^2) \to 0$.*

Proof. From the definition of α we get

$$2\alpha = \frac{1}{C_0} + \frac{S^J(a)}{\epsilon^2}. \tag{4.4.15}$$

From the definition of β we get

$$\beta = \kappa + \frac{m_0}{C_0} + \frac{1}{\epsilon^2}\sum_{j=1}^{J} a^j y_j. \tag{4.4.16}$$

Using the equation (4.3.2) we can rewrite the term $(\sum_{j=1}^{J} a^j y_j)$ as following

$$\begin{aligned}\sum_{j=1}^{J} a^j y_j &= \sum_{j=1}^{J} a^j (v_0 a^j + \nu_j) \\ &= v_0 S^J(a) + \sum_{j=1}^{J} a^j \nu_j \end{aligned} \tag{4.4.17}$$

which in turn gives

$$\beta = \kappa + \frac{m_0}{C_0} + \frac{1}{\epsilon^2}(v_0 S^J(a) + \sum_{j=1}^{J} a^j \nu_j). \tag{4.4.18}$$

Now substituting the values of α and $\left(\sum_{j=1}^{J} a^j y_j\right)$ in the equation (4.4.13) yields

$$\mathbb{E}_P[u_0] = \frac{\frac{m_0}{C_0} + \frac{v_0 S^J(a) + \sum_{j=1}^{J} a^j \nu_j}{\epsilon^2}}{\frac{1}{C_0} + \frac{1}{\epsilon^2} S^J(a)}. \tag{4.4.19}$$

On simplifying further we get the mean of posterior distribution for given realization of observation noise as

$$\mathbb{E}_P[u_0] = \frac{m_0\epsilon^2 + v_0 S^J(a) C_0 + C_0 \sum_{j=1}^{J} a^j \nu_j}{\epsilon^2 + C_0 S^J(a)} \tag{4.4.20}$$

and the variance is given by the equation (4.4.14) as

$$\begin{aligned} \mathbb{E}_P[u_0^2] - \mathbb{E}_P[u_0]^2 &= \frac{1}{2\alpha} \\ &= \frac{C_0\epsilon^2}{\epsilon^2 + C_0 S^J(a)}. \end{aligned} \tag{4.4.21}$$

Now we consider the realization of observational noise as a random variable $S^J := \sum_{j=1}^{J} a^j \nu_j$ and the mean $\mathbb{E}_P[u_0]$ can be rewritten as

$$\mathbb{E}_P[u_0] = v_0 \left(\frac{\frac{m_0\epsilon^2}{v_0 S^J(a) C_0} + 1 + \frac{S^J}{v_0 S^J(a)}}{\frac{\epsilon^2}{C_0 S^J(a)} + 1} \right).$$

Since $a > 1$ as $J \to \infty$, $S^J(a) \to \infty$ and Lemma 4.4.1 give

$$\lim_{J\to\infty} \mathbb{E}_P[u_0] \to v_0$$

almost surely. Also since $\lim_{J\to\infty} S^J(a) = \infty$ we get $\lim_{J\to\infty} (\mathbb{E}_P[u_0^2] - \mathbb{E}_P[u_0]^2) \to 0$. □

The preceding result complements Theorem 4.3.2 where in the case when $a > 1$ we showed the mean square convergence of the estimate to the true initial condition.

4.4.2 3DVAR constraint 4DVAR

In this section we look at the posterior distribution of the initial condition u_0 of the linear model described by the equations (4.3.3) and (4.3.2), given the data $\{y_j\}_{j=1}^J$. We again assume the prior distribution for the initial condition u_0 to be Gaussian given as $N(m_0, C_0)$ and the likelihood

function for the observed data can be given as

$$\mathcal{L}(\{y_j\}_{j=1}^J|u_0) = e^{-\frac{1}{2}\sum_{j=1}^J ||y_j-u_j||^2_{\epsilon^2}}$$

where $\epsilon^2 > 0$. Now we can write the posterior distribution as

$$P(u_0|\{y_j\}_{j=1}^J) = \frac{1}{Z}e^{-\frac{1}{2}||u_0-m_0||^2_{C_0}}e^{-\frac{1}{2}\sum_{j=1}^J ||y_j-u_j||^2_{\epsilon^2}}$$

where Z is normalizing constant and u_j follows the dynamics given by the equation (4.3.3) . The cumulant generating function for the above posterior distribution can be written as following

$$e^{C(\kappa)} = \frac{1}{Z}\int e^{\kappa u_0}e^{-\frac{1}{2}||u_0-m_0||^2_{C_0}}e^{-\frac{1}{2}\sum_{j=1}^J ||y_j-u_j||^2_{\epsilon^2}}\mathrm{d}u_0. \tag{4.4.22}$$

Since u_j is given by

$$u_j = a^j(1-g)^j(u_0-v_0) + a^j v_0 + \sum_{k=1}^j a^{j-k}(1-g)^{j-k}g\nu_k$$

we can expand the terms and collect coefficients of the powers of u_0 to get

$$\begin{aligned} e^{C(\kappa)} &= \frac{1}{Z}\int e^{\kappa u_0}e^{-\frac{1}{2}||u_0-m_0||^2_{C_0}}e^{-\frac{1}{2}\sum_{j=1}^J ||y_j-a^j(1-g)^j(u_0-v_0)-a^jv_0-\sum_{k=1}^j a^{j-k}(1-g)^{j-k}g\nu_k||^2_{\epsilon^2}}\mathrm{d}u_0 \\ &= \frac{1}{Z}\int e^{\kappa u_0}e^{-\frac{1}{2}||u_0-m_0||^2_{C_0}}e^{-\frac{1}{2}\sum_{j=1}^J ||\nu_j-a^j(1-g)^j(u_0-v_0)-\sum_{k=1}^j a^{j-k}(1-g)^{j-k}g\nu_k||^2_{\epsilon^2}}\mathrm{d}u_0 \\ &= \frac{1}{Z}\int e^{-\alpha u_0^2+\beta u_0+c}\mathrm{d}u_0 \end{aligned} \tag{4.4.23}$$

where α, β and c are defined as

$$\alpha = \frac{1}{2C_0} + \frac{1}{2\epsilon^2}\sum_{j=1}^J a^{2j}(1-g)^{2j} \tag{4.4.24}$$

$$\beta = \kappa + \frac{m_0}{C_0} + \frac{1}{\epsilon^2}\sum_{j=1}^J\big(a^{2j}(1-g)^{2j}v_0 + a^j(1-g)^j\nu_j - a^j(1-g)^j\sum_{k=0}^{j-1}a^k(1-g)^k\nu_{k+1}\big) \tag{4.4.25}$$

$$c = -\frac{m_0^2}{2C_0} - \frac{1}{2\epsilon^2}\sum_{j=1}^J\big(\nu_j + a^j(1-g)^jv_0 - \sum_{k=1}^j a^{j-k}(1-g)^{j-k}g\nu_k\big)^2. \tag{4.4.26}$$

Since C is the cumulant generating function for the posterior distribution we get the following relations

$$\mathbb{E}_P[u_0] = \left.\frac{\partial C}{\partial \kappa}\right|_{\kappa=0} \tag{4.4.27}$$

$$\mathbb{E}_P[u_0^2] - \mathbb{E}_P[u_0]^2 = \left.\frac{\partial^2 C}{\partial \kappa^2}\right|_{\kappa=0}. \tag{4.4.28}$$

To calculate these values we take natural Logarithm of both sides in equation (4.4.23)

$$\ln e^{C(\kappa)} = \ln\left(e^{\frac{\beta^2}{4\alpha}+c}\int e^{-(\sqrt{\alpha}u_0-\frac{\beta}{2\sqrt{\alpha}})^2}\mathrm{d}u_0\right) \tag{4.4.29}$$

$$C(\kappa) = \frac{\beta^2}{4\alpha}+c = \frac{1}{4\alpha}\left(\kappa^2+\frac{2\kappa m_0}{C_0}+\frac{2\kappa}{\epsilon^2}\sum_{j=1}^{J}\left(\tilde{a}^{2j}v_0-\sum_{k=1}^{j}\tilde{a}^{2j-k}g\nu_k+\tilde{a}^j\nu_j\right)\right)\underbrace{-\frac{m_0^2}{2C_0}-\frac{1}{\epsilon^2}\sum_{j=1}^{J}y_j^2}_{\text{terms independent of }\kappa}. \tag{4.4.30}$$

On differentiating the cumulant generating function $C(\kappa)$ we get

$$\left.\frac{\partial C}{\partial \kappa}\right|_{\kappa=0} = \frac{1}{2\alpha}\left(\frac{m_0}{C_0}+\frac{1}{\epsilon^2}\sum_{j=1}^{J}\left(\tilde{a}^{2j}v_0-\sum_{k=1}^{j}\tilde{a}^{2j-k}g\nu_k+\tilde{a}^j\nu_j\right)\right) \tag{4.4.31}$$

and differentiating again gives

$$\left.\frac{\partial^2 C}{\partial \kappa^2}\right|_{\kappa=0} = \frac{1}{4\alpha}\frac{\partial^2\beta^2}{\partial\kappa^2} = \frac{1}{2\alpha}. \tag{4.4.32}$$

Substituting these values in the equations (4.4.27) and (4.4.28) gives

$$\mathbb{E}_P[u_0] = \frac{1}{2\alpha}\left(\frac{m_0}{C_0}+\frac{1}{\epsilon^2}\sum_{j=1}^{J}\left(\tilde{a}^{2j}v_0-\sum_{k=1}^{j}\tilde{a}^{2j-k}g\nu_k+\tilde{a}^j\nu_j\right)\right), \tag{4.4.33}$$

$$\mathbb{E}_P[u_0^2] - \mathbb{E}_P[u_0]^2 = \frac{1}{2\alpha}. \tag{4.4.34}$$

Now we can give the main result of this section.

Theorem 4.4.3. *Let v be a solution of the linear system given by the equation* (4.3.1) *with initial condition v_0 and u_0 be the maximum a posteriori estimate for the distribution given by the equation* (4.4.22)*, with observations as defined in the equation* (4.3.2)*. We observe that if $\tilde{a} > 1$ then*

$$\lim_{J\to\infty} Var(\mathbb{E}_P[u_0]) \to \frac{\epsilon^2\sigma^4}{(\tilde{a}^2-1)(\sigma^2+\epsilon^2)^2}.$$

Proof. From the definition of α we get

$$2\alpha = \frac{1}{C_0}+\frac{S^J(\tilde{a})}{\epsilon^2}. \tag{4.4.35}$$

From the definition of β we get

$$\beta = \kappa + \frac{m_0}{C_0} + \frac{1}{\epsilon^2}\sum_{j=1}^{J}\big(\tilde{a}^{2j}v_0 + \tilde{a}^j\nu_j - \tilde{a}^j\sum_{k=0}^{j-1}\tilde{a}^k g\nu_{k+1}\big). \tag{4.4.36}$$

which in turn can be simplified as

$$\begin{aligned}
\beta &= \kappa + \frac{m_0}{C_0} + \frac{1}{\epsilon^2}\Big(v_0 S^J(\tilde{a}) + \sum_{j=1}^{J}\big(\tilde{a}^j\nu_j - \tilde{a}^j\sum_{k=0}^{j-1}\tilde{a}^k g\nu_{k+1}\big)\Big) \\
&= \kappa + \frac{m_0}{C_0} + \frac{1}{\epsilon^2}\Big(v_0 S^J(\tilde{a}) + \sum_{j=1}^{J}\tilde{a}^j\nu_j - \sum_{j=1}^{J}\sum_{k=1}^{j}\tilde{a}^{2j-k}g\nu_k\Big) \\
&= \kappa + \frac{m_0}{C_0} + \frac{1}{\epsilon^2}\Big(v_0 S^J(\tilde{a}) + \sum_{j=1}^{J}\tilde{a}^j\nu_j - \sum_{k=1}^{J}g\nu_k\sum_{j=k}^{J}\tilde{a}^{2j-k}\Big) \\
&= \kappa + \frac{m_0}{C_0} + \frac{1}{\epsilon^2}\Big(v_0 S^J(\tilde{a}) + \sum_{j=1}^{J}\tilde{a}^j\nu_j - \sum_{k=1}^{J}g\nu_k\frac{\tilde{a}^k - \tilde{a}^{2(J+1)-k}}{\tilde{a}^2-1}\Big)
\end{aligned} \tag{4.4.37}$$

Substituting the values of α in the equation (4.4.33) yields the mean value as

$$\mathbb{E}_P[u_0] = \Big(\frac{m_0}{C_0} + \frac{1}{\epsilon^2}\big(v_0 S^J(\tilde{a}) + \sum_{j=1}^{J}\tilde{a}^j\nu_j - \sum_{k=1}^{J}g\nu_k\frac{\tilde{a}^k - \tilde{a}^{2(J+1)-k}}{\tilde{a}^2-1}\big)\Big)\Big(\frac{1}{C_0} + \frac{1}{\epsilon^2}S^J(\tilde{a})\Big)^{-1}, \tag{4.4.38}$$

and the variance is given by the equation (4.4.34) as

$$\begin{aligned}
\mathbb{E}_P[u_0^2] - \mathbb{E}_P[u_0]^2 &= \frac{1}{2\alpha} \\
&= \frac{C_0\epsilon^2}{\epsilon^2 + C_0 S^J(\tilde{a})}.
\end{aligned} \tag{4.4.39}$$

For the convenience of notation we again consider the random variable S^J as the noise realization defined in the Lemma 4.4.1 but with parameter $\tilde{a}$ i.e. $S^J \sim N\Big(0, \epsilon^2\sum_{j=1}^{J}\tilde{a}^{2j}\Big)$, then on simplifying further we get

$$\begin{aligned}
\mathbb{E}_P[u_0] &= \frac{m_0\epsilon^2 + C_0 v_0 S^J(\tilde{a})}{\epsilon^2 + C_0 S^J(\tilde{a})} + \frac{C_0 S^J}{\epsilon^2 + C_0 S^J(\tilde{a})}\Big(1 - \frac{g}{\tilde{a}^2-1}\Big) + \frac{gC_0\sum_{k=1}^{J}\tilde{a}^{2(J+1)-k}\nu_k}{(\tilde{a}^2-1)(\epsilon^2 + C_0 S^J(\tilde{a}))} \\
&= \frac{m_0\epsilon^2 + C_0 v_0 S^J(\tilde{a})}{\epsilon^2 + C_0 S^J(\tilde{a})} + \frac{C_0 S^J}{\epsilon^2 + C_0 S^J(\tilde{a})}\Big(1 - \frac{g}{\tilde{a}^2-1}\Big) + \frac{g\tilde{a}^{J+1}C_0\sum_{k=1}^{J}\tilde{a}^k\nu_{J-k+1}}{(\tilde{a}^2-1)(\epsilon^2 + C_0 S^J(\tilde{a}))} \\
&= \frac{m_0\epsilon^2 + C_0 v_0 S^J(\tilde{a})}{\epsilon^2 + C_0 S^J(\tilde{a})} + \frac{C_0 S^J}{\epsilon^2 + C_0 S^J(\tilde{a})}\Big(1 - \frac{g}{\tilde{a}^2-1}\Big) + \frac{g\tilde{a}^{J+1}C_0 S^J}{(\tilde{a}^2-1)(\epsilon^2 + C_0 S^J(\tilde{a}))}.
\end{aligned}$$

Since $\tilde{a} > 1$ as $J \to \infty$, $S^J(a) \to \infty$ and from Lemma 4.4.1 we know the random variable $\frac{S^J}{S^J(\tilde{a})} \to 0$ almost surely hence

$$\underbrace{\mathbb{E}_P[u_0]}_{\text{when } J \to \infty} = \underbrace{\frac{C_0 v_0 S^J(\tilde{a})}{\epsilon^2 + C_0 S^J(\tilde{a})}}_{\to v_0} + \underbrace{\frac{m_0 \epsilon^2}{\epsilon^2 + C_0 S^J(\tilde{a})}}_{\to 0} + \underbrace{\frac{C_0 S^J}{\epsilon^2 + C_0 S^J(\tilde{a})}\Big(1 - \frac{g}{\tilde{a}^2 - 1}\Big)}_{\to 0 \text{ a.s.}} + \frac{g\tilde{a}^{J+1} C_0 S^J}{(\tilde{a}^2 - 1)(\epsilon^2 + C_0 S^J(\tilde{a}))}.$$

Notice that the final term in the above expression is a Gaussian random variable with zero mean and variance $\frac{g^2 \tilde{a}^{2(J+1)} C_0^2 \epsilon^2 S^J(\tilde{a})}{(\tilde{a}^2-1)^2(\epsilon^2+C_0 S^J(\tilde{a}))^2}$ for fixed J. Taking the limit $J \to \infty$ we see that

$$\lim_{J\to\infty} \frac{g^2 \tilde{a}^{2(J+1)} C_0^2 \epsilon^2 S^J(\tilde{a})}{(\tilde{a}^2 - 1)^2(\epsilon^2 + C_0 S^J(\tilde{a}))^2} = \frac{\epsilon^2 \sigma^4}{(\tilde{a}^2 - 1)(\sigma^2 + \epsilon^2)^2}$$

which is the mean squared error observed in Theorem 4.3.3.

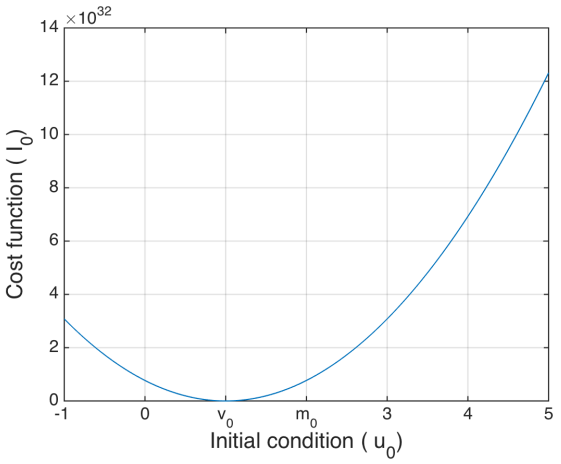
□

4.5 Numerical Results

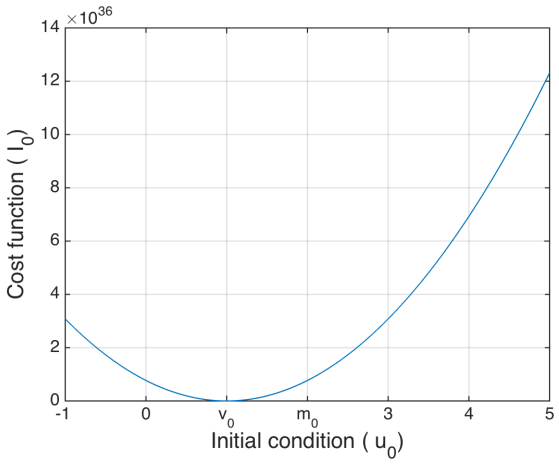
In this section we present numerical experiments in order to demonstrate the relative merits and shortcomings of the 3DVAR constraint 4DVAR method compared to Standard 4DVAR methodology in context of the linear models presented in Section 4.3. Theorems 4.3.1 and 4.3.3 express the bias in the estimation of initial condition as a function of the prior mean (m_0), the prior variance (σ^2) and the observational error variance (ϵ^2). In the following we illustrate various ways in which (σ^2) and (ϵ^2) influence the bias for linear system and in further subsections we extend the numerical experiments to nonlinear systems namely Lorenz' 63 and Lorenz' 96 models for qualitative behaviour predictions.

4.5.1 Linear System

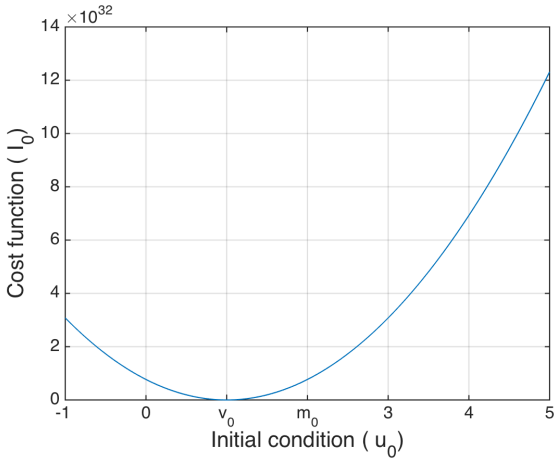
We first observe the application of standard 4DVAR scheme to the linear one dimensional system given by the equation 4.3.1 for the growth coefficients $a = 1.2$ and 0.9 in Figures 4.5.1 and 4.5.2 respectively where we plot the 4DVAR cost functional against a range of initial conditions. The true underlying initial condition is chosen to be $v_0 = 1$. The prior distribution for the initial condition is $N(m_0, \sigma^2)$ with $m_0 = 2$ and $\sigma^2 = 0.1$. The observations are made as in the equation 4.3.2 where the observational error terms ν_j are i.i.d. random variables for all j distributed as $\nu_j \sim N(0, \epsilon^2)$ with $\epsilon^2 = 1$ and the number of observations to be $J = 200$. Along with this set of parameters ($\sigma^2 = 0.1$ and $\epsilon^2 = 1$), we also look at the case when the observational noise is small ($\sigma^2 = 0.1$ and $\epsilon^2 = 1 \times 10^{-4}$) and when the prior variance is small ($\sigma^2 = 1 \times 10^{-4}$ and $\epsilon^2 = 1$).



(a) $\sigma^2 = 0.1$ and $\epsilon^2 = 1$



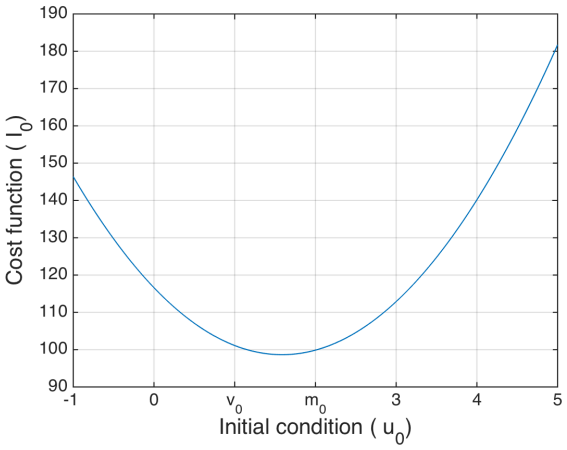
(b) $\sigma^2 = 0.1$ and $\epsilon^2 = 1 \times 10^{-4}$



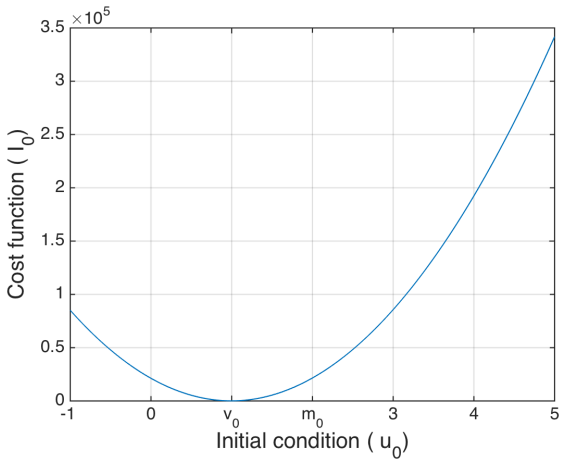
(c) $\sigma^2 = 1 \times 10^{-4}$ and $\epsilon^2 = 1$

Figure 4.5.1: The growth coefficient $a = 1.2$. In this case we see the 4DVAR cost function has a minimum at the true underlying initial condition $v_0 = 1$ for different sets of parameter values of σ^2 and ϵ^2. Identical profile in all the cases suggest that when the linear growth coefficient $a > 1$ the system retains enough information over the noisy observations to track the true initial condition irrespective of the variance in prior distribution.

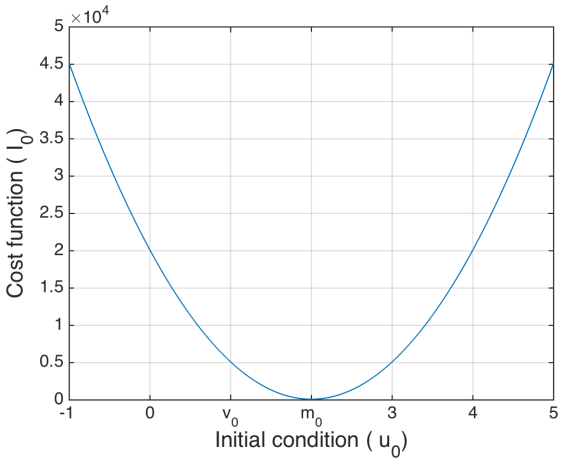
In the case when the growth coefficient $a = 0.9$ in the equation 4.3.1 we observe in Figure 4.5.2 the cost function exhibits properties as described in Remark 4.3.2.



(a) $\sigma^2 = 0.1$ and $\epsilon^2 = 1$



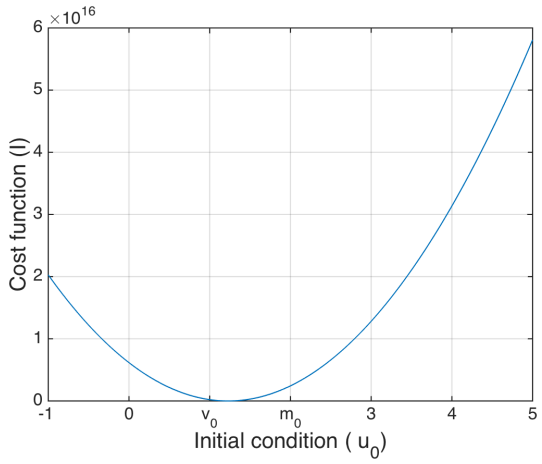
(b) $\sigma^2 = 0.1$ and $\epsilon^2 = 1 \times 10^{-4}$



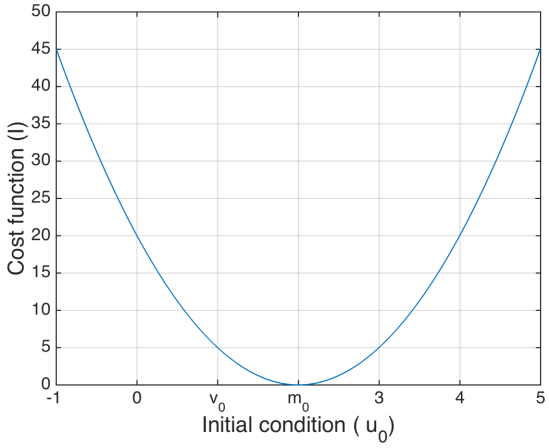
(c) $\sigma^2 = 1 \times 10^{-4}$ and $\epsilon^2 = 1$

Figure 4.5.2: The growth coefficient $a = 0.9$. In Figure 4.5.2a the minimum is obtained at $u_0 = 1.62$ which is in accordance with the bias term given in Theorem 4.3.1 which for given parameter values is $|u_0 - v_0| = 0.7162$. When the observation noise is made small ($\epsilon^2 = 10^{-4}$) the bias term given in Theorem 4.3.1 becomes small and the minimum occurs at the true initial condition v_0 (Figure 4.5.2b). On the other hand when the prior distribution variance is chosen to be small ($\sigma^2 = 10^{-4}$) the term involving prior mean m_0 in equation (4.2.4) dominates and the minimum occurs at the prior mean $m_0 = 2$ (Figure 4.5.2c).

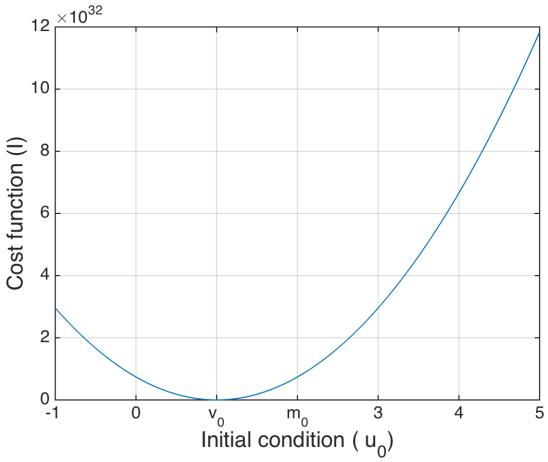
Now we look at the 3DVAR constraint 4DVAR algorithm. The 3DVAR constraint is given by the equation (4.3.3). We first choose the growth coefficient $a = 1.2$ keeping $v_0 = 1$ and $m_0 = 2$.



(a) $\sigma^2 = 0.1$ and $\epsilon^2 = 1$



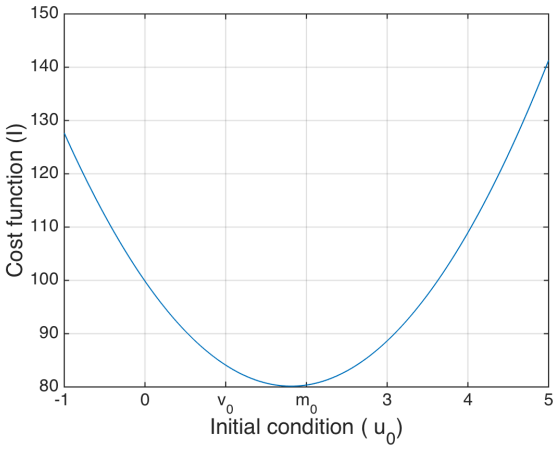
(b) $\sigma^2 = 0.1$ and $\epsilon^2 = 1 \times 10^{-4}$



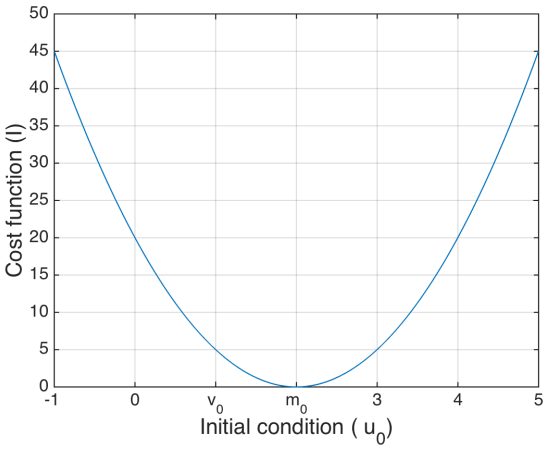
(c) $\sigma^2 = 1 \times 10^{-4}$ and $\epsilon^2 = 1$

Figure 4.5.3: The growth coefficient $a = 1.2$. In the case when $\sigma^2 = 0.1$ and $\epsilon^2 = 1$ we get the effective growth coefficient $\tilde{a} := \frac{a\epsilon^2}{\epsilon^2+\sigma^2} = 1.091$. We observe the minimum at $u_0 = 1.23$ in Figure 4.5.3a. Since $|\tilde{a}| > 1$ Theorem 4.3.3 gives us the bias value $|u_0 - v_0| = 0.2085$. For small observational noise i.e. $\epsilon^2 = 10^{-4}$ the effective growth coefficient becomes $\tilde{a} = 1.199 \times 10^{-3}$. In Figure 4.5.3b we observe the minimum is at $u_0 = m_0 = 2$ which agrees with the Remark 4.3.4. Finally for small prior variance $\sigma^2 = 10^{-4}$ the effective growth coefficient is $\tilde{a} = 1.199$ and the minimum occurs at $u_0 = v_0 = 1$ as observed in Figure 4.5.3c. The profile is similar to the Figure 4.5.1c which suggests that in the 3DVAR constraint effectively ignores the innovation terms.

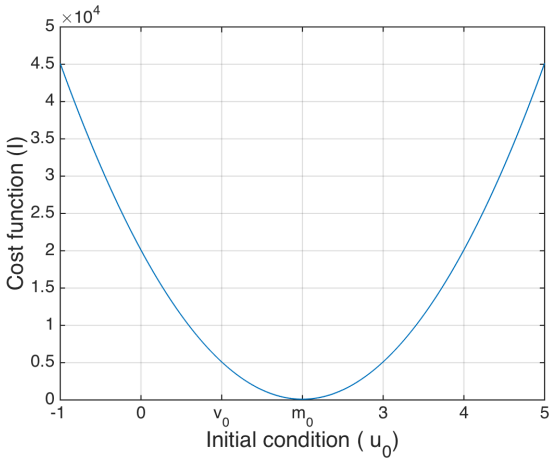
Finally we consider the application of 3DVAR constraint 4DVAR sceme to the case when the growth coefficient is chosen to be $a = 0.9$ in the equation (4.3.3).



(a) $\sigma^2 = 0.1$ and $\epsilon^2 = 1$



(b) $\sigma^2 = 0.1$ and $\epsilon^2 = 1 \times 10^{-4}$



(c) $\sigma^2 = 1 \times 10^{-4}$ and $\epsilon^2 = 1$

Figure 4.5.4: The growth coefficient $a = 0.9$. For values $\sigma^2 = 0.1$ and $\epsilon^2 = 1$ the minimum occurs at $u_0 = 1.78$ in Figure 4.5.4a. Theorem 4.3.3 informs us the bias value $|u_0 - v_0| = 0.8416$. For small observational noise again the minimum is observed at the prior mean $m_0 = 1$ in Figure 4.5.4b, as the effective growth coefficient becomes $\tilde{a} = 8.991 \times 10^{-4}$. For the case when the prior variance $\sigma^2 = 10^{-4}$ is small, the minimum occurs at the prior mean $m_0 = 2$ (Figure 4.5.4c) similar to the case for standard 4DVAR scheme (Figure 4.5.2c).

4.5.2 Nonlinear Models

In this section we focus on nonlinear chaotic models. For nonlinear chaotic models we expect the standard 4DVAR functional to provide a rough surface with multiple local minima when plotted as a function of initial conditions due to sensitive dependence of the underlying model on initial conditions. The application of 3DVAR constraint 4DVAR algorithm behaves differently depending upon the accuracy of the 3DVAR filter. The time averaged root mean squared error (RMSE) is used to evaluate the accuracy of the 3DVAR scheme. For N dimensional system with true state $v = (v^{(1)}, \ldots, v^{(N)})$, J observations steps and the filtering estimate $u = (u^{(1)}, \ldots, u^{(N)})$, RMSE can

be expressed as following

$$RMSE = \frac{1}{J}\sum_{j=1}^{J}\sqrt{\frac{1}{N}\sum_{n=1}^{N}(v_j^{(n)} - u_j^{(n)})^2}. \tag{4.5.1}$$

When the 3DVAR scheme does not track the true trajectory accurately, i.e. the RMSE does not converge to $\mathcal{O}(\epsilon)$ neighbourhood of the truth trajectory, the process fed in to the 4DVAR functional does not provide enough information about the true underlying process so we expect the minimization profile provided by the variational form I to be rugged whereas in the case when the 3DVAR scheme is accurate in tracking the true trajectory, we expect the minimization profile to be smooth.

The set up for numerical experiments in case of nonlinear models is as follows

1. Application of standard 4DVAR scheme.
2. Application of 3DVAR constraint 4DVAR scheme in following cases,
   - 3DVAR scheme does not track the true trajectory accurately,
   - 3DVAR scheme tracks the true trajectory accurately,
   - The prior variance σ^2 is large.
3. Application of 3DVAR constraint 4DVAR scheme when the observation noise is small.

Subsections 4.5.2 and 4.5.2 examine the bias in varying regimes of (σ^2) and (ϵ^2) when the underlying dynamical system is the Lorenz'63 model and Lorenz'96 model respectively.

Lorenz'63 Model

The Lorenz equations [53, 74, 22, 29] are a system of three coupled non-linear ordinary differential equations whose solution $u \in \mathbb{R}^3$, where $u = (u_x, u_y, u_z)$, satisfies

$$\frac{du}{dt} + Au + B(u,u) = f, \quad u(0) = u_0, \tag{4.5.2}$$

where

$$A = \begin{pmatrix} \alpha & -\alpha & 0 \\ \alpha & 1 & 0 \\ 0 & 0 & b \end{pmatrix}, \quad f = \begin{pmatrix} 0 \\ 0 \\ -b(r+\alpha) \end{pmatrix}$$

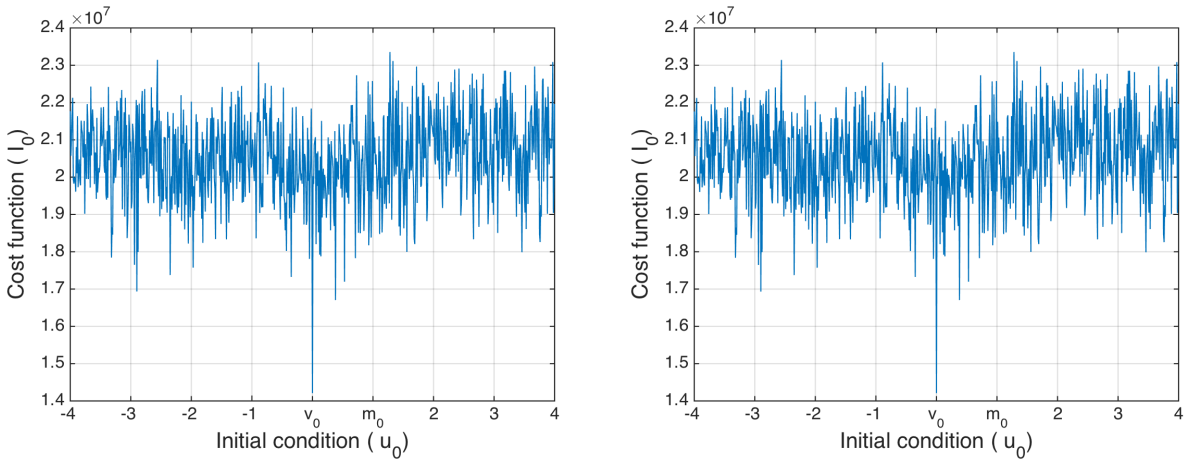
$$B(u,\tilde{u}) = \begin{pmatrix} 0 \\ (u_x\tilde{u_z} + u_z\tilde{u_x})/2 \\ -(u_x\tilde{u_y} + u_y\tilde{u_x})/2 \end{pmatrix}.$$

In our numerical experiments we use the classical parameter values $(\alpha, b, r) = (10, \frac{8}{3}, 28)$, at which the system is chaotic [79].

For our experiments we fix the initial condition for the underlying system $u(t = 0) := (u_x, u_y, u_z)(t = 0) = (0, 1, 1)$. The trajectory followed from this initial condition is henceforth referred as the truth or underlying true trajectory. For the approximation we choose prior mean $m_0 := (m_x, m_y, m_z)(t = 0) = (1, 1, 1)$, so it differs from the true initial condition only in the x-component. The prior covariance is denoted as $C_0 := \sigma^2 \mathbb{I}_{3\times 3}$. Synthetic observations are generated at the observation times by adding mean zero Gaussian noise to the underlying truth trajectory. Since the observation errors are assumed to be uncorrelated, the observation error covariance matrix $\Gamma := \epsilon^2 \mathbb{I}_{3\times 3}$ is diagonal.

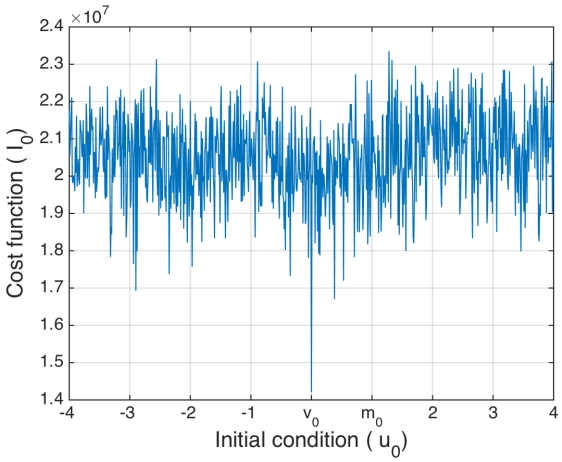
We observe all the components of the three dimensional system in observation intervals of $h = 0.1$, i.e., $t_j = 0.1j$, and with observation error variance $\epsilon^2 = 0.01$. A total of $J = 1000$ assimilation steps are performed. The differential equations are solved numerically with a step-size of $\triangle t = 0.01$. We carry out experiments for standard 4DVAR and 3DVAR constraint 4DVAR schemes where the respective cost functions are plotted against a range of values for the x-component of the initial condition.

When seen as a 1-dimensional minimization problem the minimization profile, given by the standard 4DVAR variational form I against various initial values of x-component, is rugged due to the chaotic nature of the underlying dynamical system. It possesses clear minimum at the true initial value $u_x(t = 0) = 0$ but has multiple local minimum as observed in Figure 4.5.5 for different values of prior distribution variance σ^2. This feature renders derivative based minimization schemes ineffective and sampling based schemes overly expensive.



(a) $\sigma^2 = 9 \times 10^{-4}$

(b) $\sigma^2 = 1.6 \times 10^{-3}$



(c) $\sigma^2 = 1 \times 10^{-2}$

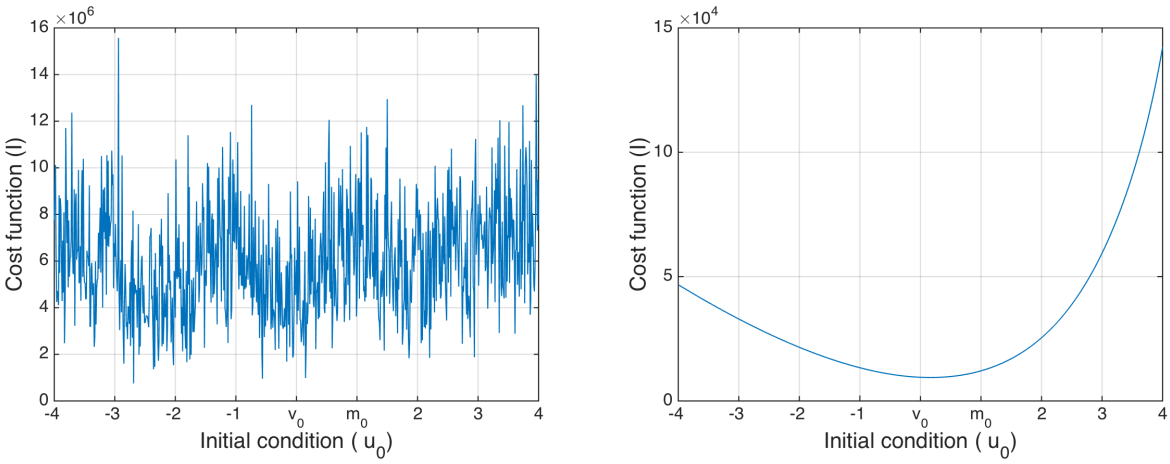
Figure 4.5.5: Standard 4DVAR cost function profile for Lorenz'63 system for observational noise variance $\epsilon^2 = 1 \times 10^{-2}$.

In contrast when we plot the minimization profile, given by the variational form I constrained by the 3DVAR algorithm i.e. the approximation process follows the dynamics described by the 3DVAR algorithm (4.2.5) rather than the original dynamics we see that the minimization profile exhibits variable behaviour for different values of the prior variance σ^2 in regards to the salient features of the location of the minimum and the roughness of the profile.

When the value of prior variance is chosen to be $\sigma^2 = 9 \times 10^{-4}$ we observe that the approximation process given by 3DVAR scheme fails to accurately track the underlying system trajectory and the observation error term, given as $\frac{1}{2\epsilon^2} \sum_{j=1}^{J} ||y_j - u_j||^2$ in the definition of the variational function I, remains significant which results in the rugged minimization surface no clear global minimum as seen in the Figure 4.5.6a. Lack of clear minimum in contrast with the case of standard 4DVAR minimization can be attributed to the fact that the approximation process is not constrained by the underlying model dynamics but by the 3DVAR filtering scheme.

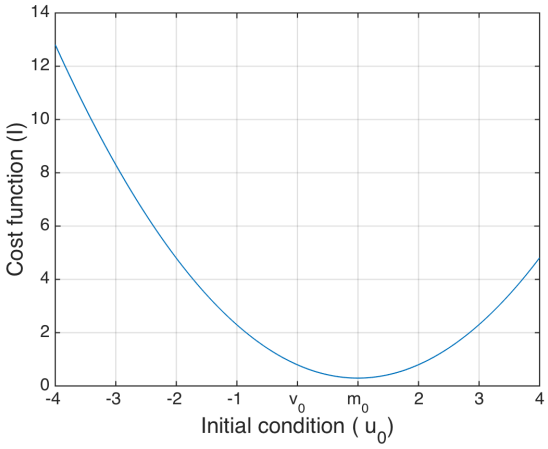
However when we increase the value of prior distribution variance to be $\sigma^2 = 1.6 \times 10^{-3}$ the approximation process given by 3DVAR scheme tracks the true trajectory accurately over time reducing the contribution of observation error term the variational function I. This results in smoothening of the minimization surface as reflected in the Figure 4.5.6b. However the minimum is observed at the initial condition $u_0 = 0.16$ instead of the underlying true initial condition $v_0 = 0$. This result indicates that if curated carefully for the bias expected in the minimum, application of 3DVAR constraint 4DVAR scheme can significantly reduce the difficulties in minimization even when the underlying system has sensitive dependence on initial condition.

On further increasing the prior variance value $\sigma^2 = 1 \times 10^{-2}$ minimum shifts to the prior mean $m_x(0) = 1$ which corresponds to the result in the Remark 4.3.4 where we observed the inverse relationship between σ^2 and the growth coefficient $\tilde{a}$ and when σ^2 is chosen large, $\tilde{a}$ becomes small and Theorem 4.3.3, although not directly applicable to nonlinear systems, suggests that minimum will occur at the prior mean as seen in the Figure 4.5.6c.



(a) $\sigma^2 = 9 \times 10^{-4}$

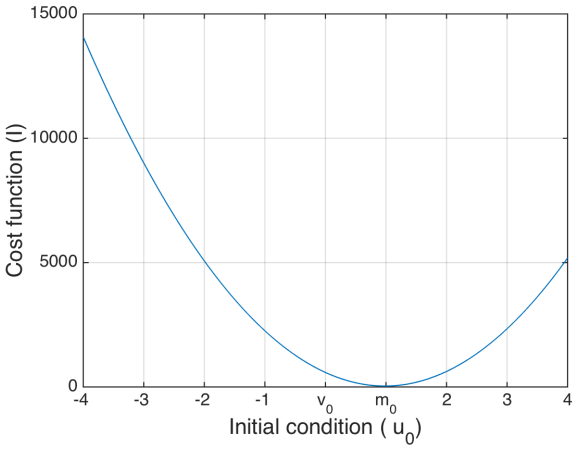
(b) $\sigma^2 = 1.6 \times 10^{-3}$



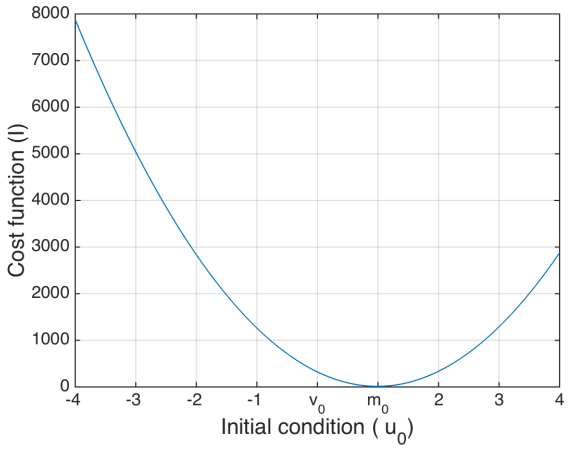
(c) $\sigma^2 = 1$

Figure 4.5.6: 3DVAR constraint 4DVAR cost function profile for Lorenz'63 system (a) when the 3DVAR filter fails to track the true underlying process, (b) when the 3DVAR filter is accurate and (c) when prior variance value is chosen to be comparatively large.

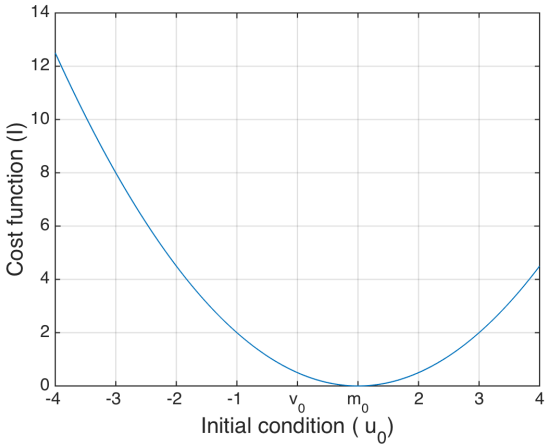
Furthermore one can consider the case when the observational noise is small. In Figure 4.5.7 we notice that when the observational noise variance $\epsilon^2 = 1 \times 10^{-4}$ is taken to be small the minimum is observed at the prior mean $m_x(0) = 1$ irrespective of the value of the prior variance σ^2 which is again in accordance with the Remark 4.3.4.



(a) $\sigma^2 = 9 \times 10^{-4}$ and $\epsilon^2 = 1 \times 10^{-4}$



(b) $\sigma^2 = 1.6 \times 10^{-3}$ and $\epsilon^2 = 1 \times 10^{-4}$



(c) $\sigma^2 = 1$ and $\epsilon^2 = 1 \times 10^{-4}$

Figure 4.5.7: 3DVAR constraint 4DVAR cost function profile for Lorenz'63 system when the observational noise is chosen to be $\epsilon^2 = 1 \times 10^{-4}$.

Lorenz'96 Model

The Lorenz 96 system as introduced in [54] is a commonly used nonlinear model for testing data assimilation schemes. The Lorenz 96 system involves a set of N variables $u = (u^{(1)}, \ldots, u^{(N)})^T \in \mathbb{R}^N$ which satisfy the following coupled ODE's

$$\frac{\mathrm{d}u^{(n)}}{\mathrm{d}t} = u^{(n-1)}(u^{(n+1)} - u^{(n-2)}) - u^{(n)} + F \quad \text{for } n = 1, 2, \ldots, N, \tag{4.5.3}$$

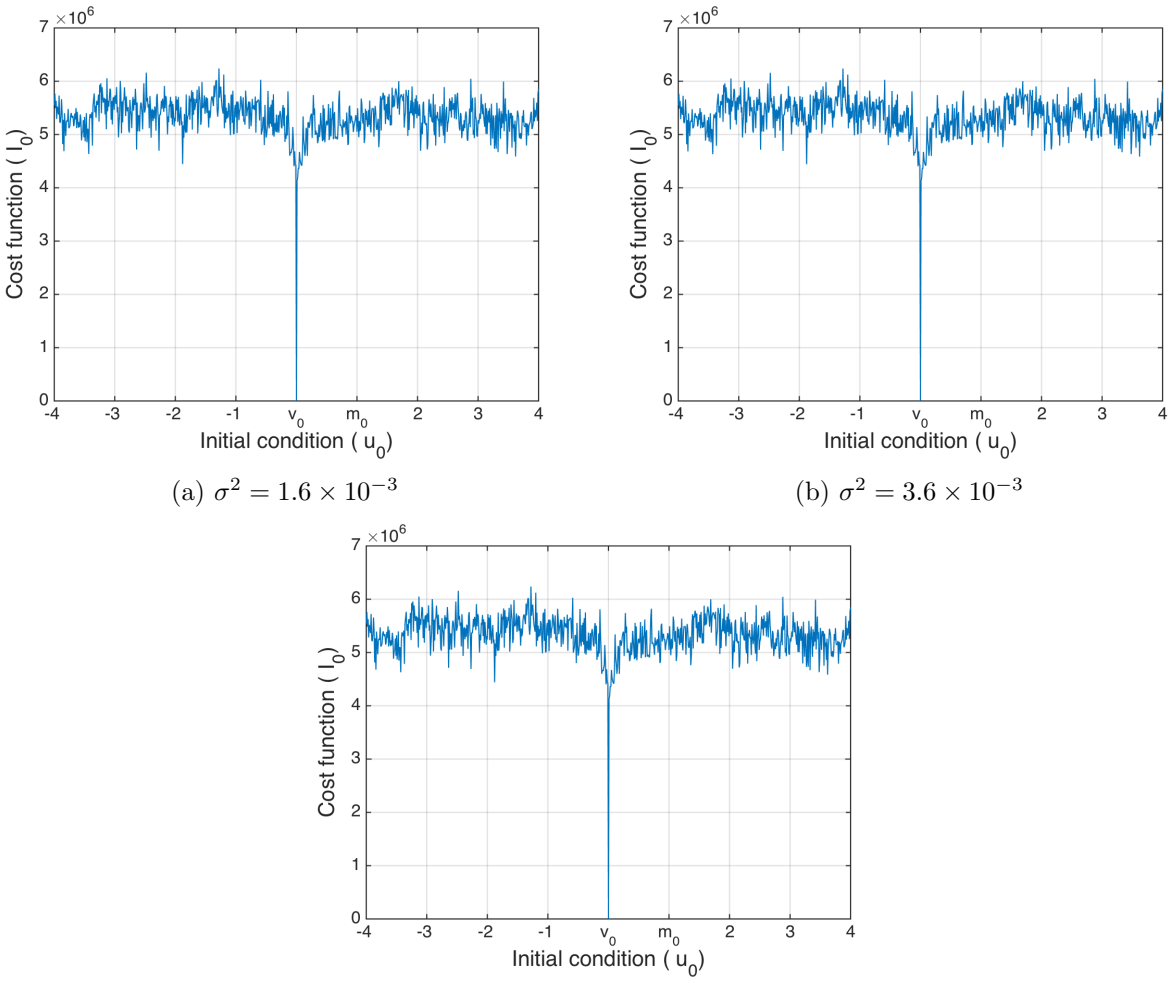
subject to the periodic boundary conditions $u^{(n-N)} = u^{(n+N)} = u^{(n)}$. We choose the dimension of the system $N = 40$ and the forcing parameter $F = 8$. For these values the system exhibits chaotic behaviour [55].

Similar to the case for Lorenz' 63 system we fix the initial condition for the underlying

Lorenz' 96 system and generate the underlying truth trajectory following the Lorenz'96 model dynamics. The prior mean is chosen so that it differs from the true initial condition only in the first component ($u^{(1)}$) and the prior covariance is defined as $C_0 := \sigma^2 \mathbb{I}_{N\times N}$. We further generate the observations by adding zero mean, uncorrelated Gaussian noise with error covariance matrix $\Gamma := \epsilon^2 \mathbb{I}_{N\times N}$ to the underlying truth trajectory.

As previously the frequency between observations is again chosen to be $h = 0.1$ which sets $t_j = 0.1j$. A total of $J = 1000$ assimilation steps are performed. The differential equations are solved numerically with a step-size of $\triangle t = 0.01$.

In this case we plot the standard 4DVAR cost functional against a range of first component ($u^{(1)}$) values of initial condition. Due to the chaotic nature of Lorenz'96 system for chosen parameter values we expect rugged minimization profile. Figure 4.5.8 shows that for standard 4DVAR variational form we get the global minimum near the true initial condition v_0 independently of the variance of prior distribution. However the variational surface is rugged and contains multiple local minima. Presence of global minimum near the true initial condition is also notable.



(a) $\sigma^2 = 1.6 \times 10^{-3}$

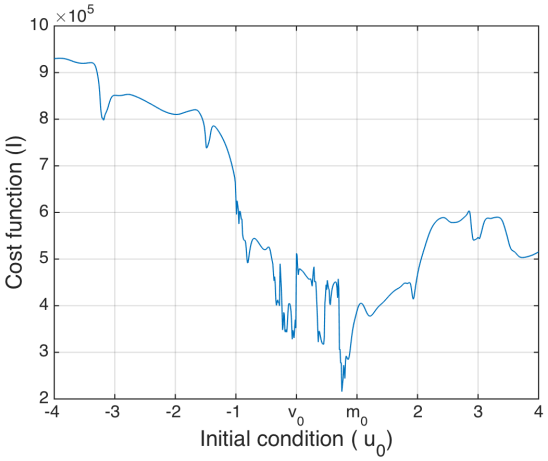
(b) $\sigma^2 = 3.6 \times 10^{-3}$

(c) $\sigma^2 = 1$

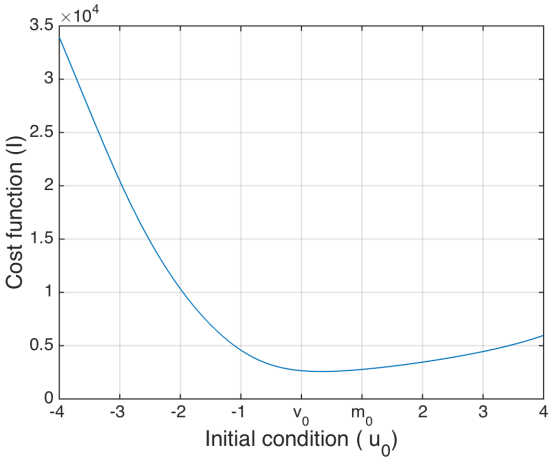
Figure 4.5.8: Standard 4DVAR cost function profile for Lorenz'96 system for observational noise variance $\epsilon^2 = 1 \times 10^{-2}$.

When the approximation process follows the 3DVAR dynamics instead of the underlying dynamics we again observe distinct appearance of the cost function depending upon the accuracy of the 3DVAR process in tracking the true underlying process. For 3DVAR constraint 4DVAR algorithm we see that when the prior variance is chosen to be $\sigma^2 = 1.6 \times 10^{-3}$ the 3DVAR process does not track the underlying process accurately which leads to the minimization profile being irregular with multiple local minima and the global minimum is not at true initial condition v_0. When the prior variance value is increased to be $\sigma^2 = 3.6 \times 10^{-3}$ the accuracy of the 3DVAR scheme increases and the minimization profile becomes smooth as in Figure 4.5.9b, however the minimum is observed at $u_0 := u^{(1)}(t = 0) = 0.33$ which leads to the presence of bias if this minimization profile is used to estimate the true initial condition. As we increase the value of the prior variance the minimum shifts towards the prior mean m_0. For the value $\sigma^2 = 1$ minimum is observed at the prior mean $m_0 := m^{(1)}(t = 0) = 1$. Again we see that although Theorem 4.3.3 is not directly applicable

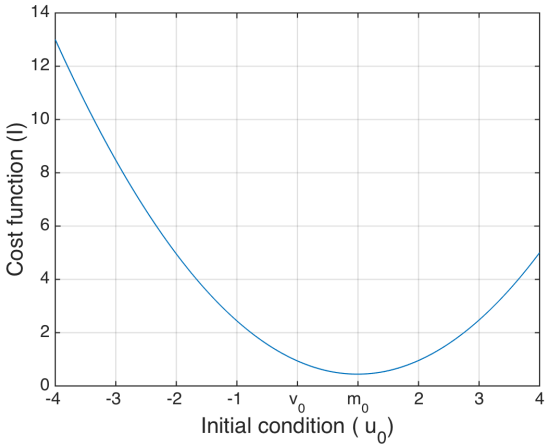
for the underlying system being nonlinear, one can extend the implications made in Remark 4.3.4 to nonlinear systems.



(a) $\sigma^2 = 1.6 \times 10^{-3}$ and $\epsilon^2 = 10^{-2}$



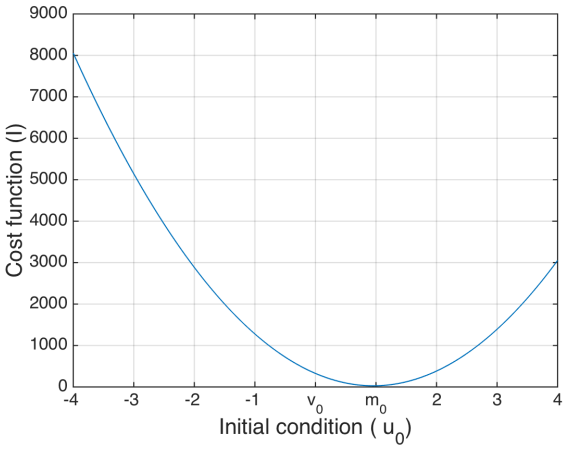
(b) $\sigma^2 = 3.6 \times 10^{-3}$ and $\epsilon^2 = 10^{-2}$



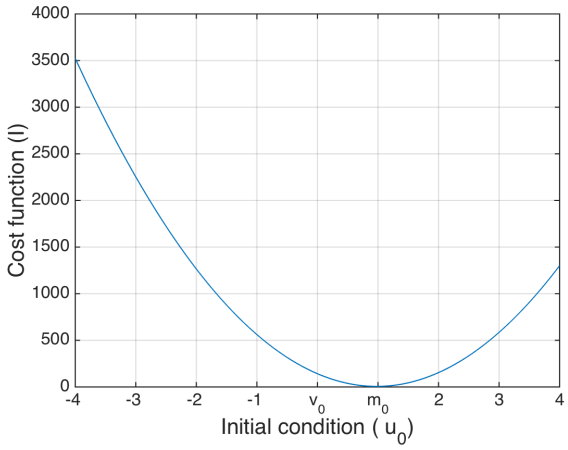
(c) $\sigma^2 = 1$ and $\epsilon^2 = 10^{-2}$

Figure 4.5.9: 3DVAR constraint 4DVAR cost function profile for Lorenz'96 system (a) when the 3DVAR filter fails to track the true underlying process, (b) when the 3DVAR filter is accurate and (c) when prior variance value is chosen to be comparatively large.

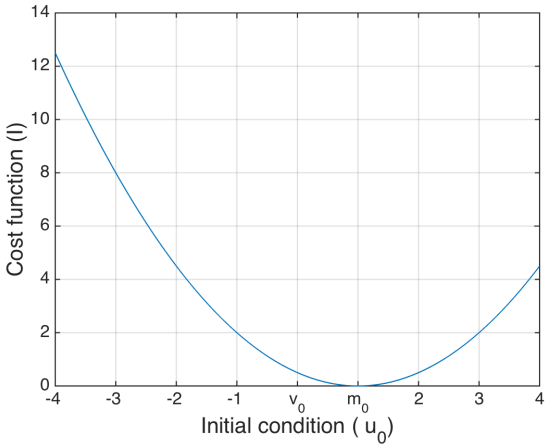
Finally for small noise case again we see that the minimization surface is smooth but the minimum is observed at the prior mean m_0 irrespective of the value of the prior variance as seen in Figure 4.5.10.



(a) $\sigma^2 = 1.6 \times 10^{-3}$ and $\epsilon^2 = 1 \times 10^{-4}$



(b) $\sigma^2 = 3.6 \times 10^{-3}$ and $\epsilon^2 = 1 \times 10^{-4}$



(c) $\sigma^2 = 1$ and $\epsilon^2 = 1 \times 10^{-4}$

Figure 4.5.10: 3DVAR constraint 4DVAR cost function profile for Lorenz'96 system when the observational noise is chosen to be $\epsilon^2 = 1 \times 10^{-4}$.

4.6 Conclusions

In this chapter we have stated and analysed a 4DVAR assimilation scheme constrained by 3DVAR estimates, applied to a simple linear state data assimilation problem. We establish the asymptotic mean square convergence of the 4DVAR estimates of the initial condition to the true initial condition for growing linear scalar systems and existence of a finite bias when the underlying linear system is contracting. We then studied the same problem under 3DVAR constraint 4DVAR scheme where we observed the existence of the bias in the initial condition for both growing and contracting linear systems. We also established the almost sure convergence of the 4DVAR estimate for initial condition using the path integral approach. Furthermore, under the path integral framework, we showed that the asymptotic variance of the estimate of initial condition from 3DVAR constraint 4DVAR formulation agrees with the bias estimate given by the mean square convergence result.

Numerical results were presented in Section 4.5.1 for the linear system and in Section 4.5.2 for Lorenz'63 and Lorenz'96 systems. The results for linear system affirm the derived analytical results. Moreover, the application of 3DVAR constraint 4DVAR to nonlinear chaotic problems accentuates the utility of this methodology. We observe that for chaotic dynamical systems, the cost function surface provided by 3DVAR constraint 4DVAR scheme is smoother hence more amenable to minimization schemes in comparison to the cost function surface provided by standard strong constraint 4DVAR scheme. However, this ease of minimization comes at the penalty of the introduction of bias in the minimum point. The main scientific challenge in applying 3DVAR constraint 4DVAR scheme is to estimate the bias for given dynamical system, which may turn out to be arduous computational problem. Nonetheless, if good estimates for the potential bias are available for the system, 3DVAR constraint 4DVAR can be applied to make the minimization process efficient and computationally cheaper.

We identify that there are two main challenges in implementing the 3DVAR constrained 4DVAR scheme. The first Challenge is to find the appropriate Kalman gain factor as seen in the section 4.5.1 where 3DVAR process accurately tracks the underlying process without over-saturating the underlying signal. The second challenge is the pre-computation of the bias given the model and the filtering parameters. In the section 4.3.2 we provide expressions for computing the bias for linear system however, computing the bias for more complex non-linear models remains an open problem.

In the next chapter we extend the application of 3DVAR constraint to the weak constraint formulation of 4DVAR assimilation scheme.

Chapter 5

3DVAR constraint Weak 4DVAR Scheme

Weak 4DVAR has been proposed to address the situations when the model does not capture the underlying system entirely. Hence, even if the true initial condition is known the system trajectory can not be reproduced by integrating the model forward in time. Weak 4DVAR scheme takes the approach of estimating the whole trajectory. The model is applied as an approximate constraint using the sequence of model error variables. This approach was first introduced by Sasaki [69], and explored further by [15, 76]. Comparison between strong constraint and weak constraint can be found in [46]. In this chapter we analyse the 3DVAR constraint in the context of weak 4DVAR scheme. As in the case of strong constraint 4DVAR, the model dynamics constraint is replaced by the 3DVAR approximation process in formulation of 3DVAR constraint weak 4DVAR. In Section 5.1 we describe the weak constraint 4DVAR formulation and 3DVAR constraint 4DVAR with respective underlying statistical assumptions. In Section 5.2 we state the linear dynamical system under consideration and establish upper bounds on the error present in the estimate. Numerical experiments for the linear system are presented in Section 5.4.1 and compared with theoretical results. Numerical results for Nonlinear systems are presented in Section 5.4.2. Summary of this chapter is presented in Section 5.5.

5.1 Set up

Given that the available model does not capture the underlying system entirely a correction term is introduced [15, 32] in the model equation. The correction term models the departure of the model equations from the actual physical system. Assuming the model is not perfect the discretized model equation is modified to include correction terms ξ_j at each time step t_j, such that:

$$u_j = \Psi(u_{j-1}) + \xi_j,\ j \in \{1, \ldots, J\}. \tag{5.1.1}$$

where the ξ_j's represent the model error terms, assumed to be zero in the perfect-model case. The variable ξ_j has the same dimension as the underlying state and represents the model error at j-th step. The model error term ξ_j can represent a systematic error, introduced by inaccurate assumptions/parameters or numerical round off errors, or stochastic errors present due to unaccounted stochastic factors in the underlying physical system. In general both of these errors are present for the modelling of physical systems. In this work we assume the model error to be stochastic in nature and normally distributed with zero mean and covariance matrix Σ_j. The observations for the underlying system are given as

$$y_j = H_j v_j + \nu_j,\ j \in \{1, \ldots, J\} \tag{5.1.2}$$

with observation operator H_j and observation error ν_j distributed as $\nu_j \sim N(0, \Gamma_j)$.

The objective function can then be extended in the absence of systematic errors as follows:

$$\mathsf{I}\big(u_0, \xi\big) = \frac{1}{2}\sum_{j=1}^{J} ||y_j - H_j u_j||^2_{\Gamma_j} + \frac{1}{2}||u_0 - m_0||^2_{C_0} + \frac{1}{2}\sum_{j=0}^{J-1} ||\xi_j||^2_{\Sigma_j}. \tag{5.1.3}$$

where, as before, C_0 is the background error covariance matrix, and Γ_j and Σ_j are the observation and model error covariance matrices, respectively and variable u_j follow the equation (5.1.1). We also make the assumption that the observation error and the model error are uncorrelated with each other and do not depend upon the state of the system. Note that the control vector here (with respect to which the functional I is minimized) is the initial condition and the sequence of model error variables $\{\xi_j\}_{j=1}^{J}$ The overall size of the problem is now multiplied by the total number of time steps. For 3DVAR constraint weak 4DVAR the state variable u_j follow the equati on

$$u_j = \Psi(u_{j-1}) + \xi_j + G_j(y_j - H_j u_j),\ j \in \{1, \ldots, J\} \tag{5.1.4}$$

where G_j is Kalman Gain matrix at time j-th time step. In the further sections we will be using the following version of Lax-Milgram Theorem:

Proposition 5.1.1. *Let V be a Hilbert Space with norm $\|\cdot\|_V$ and scalar product $\langle\cdot,\cdot\rangle_V$ and assume that $B(\cdot,\cdot)$ is a bilinear functional and $L(\cdot)$ is a linear functional that satisfy*

1. *B is symmetric, i.e. $B(u,v) = B(v,u)$, $\forall u, v \in V$,*
2. *B is $V-$elliptic i.e. $\exists \alpha > 0$, such that $B(v,v) \geq \alpha \|v\|^2_V$,*
3. *B is continuous i.e. $\exists C_B \in \mathbb{R}$ such that $|B(u,v)| \leq C_B \|u\|_V \|v\|_V$, and*
4. *L is continuous i.e. $\exists C_L \in \mathbb{R}$ such that $|L(u)| \leq C_L \|u\|_V$*

Then there is a unique function $u \in V$ such that $B(u,v) = L(v)$, $\forall v \in V$, and the stability estimate $\|u\|_V \leq \frac{C_L}{\alpha}$ holds.

5.2 Standard 4DVAR with Model Error

Linear Model

We consider the application of 4DVAR smoother method to a linear one dimensional discrete dynamical system. Let $v \in \mathbb{R}^{J+1}$ be the true underlying solution with fixed $J \in \mathbb{Z}^+$. The solution satisfies the relation

$$v_j - av_{j-1} = 0, \quad \forall j \in \{1, \ldots, J\}, \tag{5.2.1}$$

where $a \in \mathbb{R}^+$ and the initial condition $v_0 \in \mathbb{R}$ is fixed. The true process v is observed noisily at each iteration. The observations of the system are denoted as $y := \{y_j\}_{j=1}^J$ and defined as

$$y_j = v_j + \nu_j, \tag{5.2.2}$$

where ν_j are i.i.d.random variables distributed as $\nu_j \sim N(0, \epsilon^2)$. We model the underlying linear one dimensional process as following

$$u_j = \Psi(u_{j-1}) := au_{j-1} + \xi_j \tag{5.2.3}$$

with the initial condition $u_0 \sim N(m_0, \sigma_0^2)$ and the model error $\xi_j \sim N(0, \sigma^2)$. Now we define the weak constraint 4DVAR minimization functional $\mathsf{I}(\cdot,\cdot) : \mathbb{R}^{J+1} \times \mathbb{R}^J \to \mathbb{R}$ as following:

$$\mathsf{I}\big(u, \xi\big) = \frac{1}{2\sigma_0^2}(u_0 - m_0)^2 + \frac{1}{2\epsilon^2}\sum_{j=1}^{J}(y_j - u_j)^2 + \frac{1}{2\sigma^2}\sum_{j=1}^{J}\xi_j^2. \tag{5.2.4}$$

Since the minimization of the cost function is constrained by the equation (5.2.3) for the state variable u so we can rewrite the minimization functional as $\mathsf{I}^u(\cdot) : \mathbb{R}^{J+1} \to \mathbb{R}$, in terms of the state variable as

$$\mathsf{I}^u\big(\{u_j\}_{j=0}^J\big) = \frac{1}{2\sigma_0^2}(u_0 - m_0)^2 + \frac{1}{2\epsilon^2}\sum_{j=1}^{J}(u_j - y_j)^2 + \frac{1}{2\sigma^2}\sum_{j=1}^{J}(u_j - au_{j-1})^2. \tag{5.2.5}$$

The state vector u which minimizes the variational form 5.2.5 is called the 4DVAR estimate solution for the underlying dynamical system. Now for the convenience of the notation we define the parameters $\gamma_0^2 = \frac{\sigma^2}{\sigma_0^2}$, $\gamma^2 = \frac{\sigma^2}{\epsilon^2}$. On simplifying and rewriting the variational form 5.2.5 we get

$$\begin{aligned}
\mathsf{I}^u\big(\{u_j\}_{j=0}^J\big) &= \frac{\sigma^2}{2\sigma_0^2}(u_0 - m_0)^2 + \frac{\sigma^2}{2\epsilon^2}\sum_{j=1}^{J}(u_j - v_j - \nu_j)^2 + \frac{1}{2}\sum_{j=1}^{J}(u_j - au_{j-1})^2 \qquad (5.2.6)\\
&= \frac{\gamma_0^2}{2}(u_0 - m_0)^2 + \frac{\gamma^2}{2}\sum_{j=1}^{J}(u_j - v_j - \xi_j)^2 + \frac{1}{2}\sum_{j=1}^{J}(u_j - au_{j-1})^2\\
&= \frac{1}{2}\mathsf{a}(u,u) - \mathsf{L}(u) + \text{Constant term},
\end{aligned}$$

where the bilinear form $\mathsf{a}(\cdot,\cdot)$ and the linear functional $\mathsf{L}(\cdot)$ are defined as follows

$$\mathsf{a}(u,\lambda) = \gamma_0^2 u_0\lambda_0 + \gamma^2\sum_{j=1}^{J} u_j\lambda_j + \sum_{j=1}^{J}(u_j - au_{j-1})(\lambda_j - a\lambda_{j-1}) \tag{5.2.7}$$

$$\mathsf{L}(u) = \gamma_0^2 m_0 u_0 + \gamma^2\sum_{j=1}^{J}(v_j + \nu_j)u_j. \tag{5.2.8}$$

Now we list the key properties of the bilinear form $\mathsf{a}(\cdot,\cdot)$ and the linear functional $\mathsf{L}(\cdot)$ which help us in deriving an estimate for the minimizer of the variational form 5.2.5.

Properties 5.2.1. *The bilinear form* $\mathsf{a}(\cdot,\cdot) : \mathbb{R}^{J+1}\times\mathbb{R}^{J+1} \to \mathbb{R}$ *as defined by* (5.2.7) *is symmetric, continuous and coercive.*

Proof. The bilinear form $\mathsf{a}(\cdot,\cdot)$ as defined in the equation (5.2.7) is clearly symmetric. For the continuity of the bilinear form, consider the following rearrangement of $\mathsf{a}(\cdot,\cdot)$ as

$$\begin{aligned}
\mathsf{a}(u,\lambda) &= \gamma_0^2 u_0\lambda_0 + \gamma^2\sum_{j=1}^{J} u_j\lambda_j + \sum_{j=1}^{J}(u_j - au_{j-1})(\lambda_j - a\lambda_{j-1}) \\
&= \gamma_0^2 u_0\lambda_0 + \gamma^2\sum_{j=1}^{J} u_j\lambda_j + \sum_{j=1}^{J}\big(u_j\lambda_j - au_{j-1}\lambda_j - au_j\lambda_{j-1} + a^2u_{j-1}\lambda_{j-1}\big) \\
&= (\gamma_0^2 + a^2)u_0\lambda_0 + (\gamma^2 + 1 + a^2)\sum_{j=1}^{J-1} u_j\lambda_j + (\gamma^2+1)u_J\lambda_J - a\sum_{j=1}^{J}\big(u_j\lambda_{j-1} + u_{j-1}\lambda_j\big) \\
&= (\gamma^2 + 1 + a^2)\sum_{j=0}^{J} u_j\lambda_j + (\gamma_0^2 - \gamma^2 - 1)u_0\lambda_0 - a^2u_J\lambda_J - a\sum_{j=1}^{J}\big(u_j\lambda_{j-1} + u_{j-1}\lambda_j\big) \\
&= u^T\mathsf{A}\lambda + (\gamma_0^2 - \gamma^2 - 1 + a)u_0\lambda_0 + (a - a^2)u_J\lambda_J
\end{aligned} \tag{5.2.9}$$

where A is a balanced tridiagonal matrix given as

$$\mathsf{A} := \begin{pmatrix}
(\gamma^2+1+a^2) - a & -a & 0 & 0 & \dots & 0 & 0 \\
-a & \gamma^2+1+a^2 & -a & 0 & \dots & 0 & 0 \\
0 & -a & \gamma^2+1+a^2 & -a & \dots & 0 & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & 0 & 0 & 0 & \dots & -a & (\gamma^2+1+a^2) - a
\end{pmatrix}_{(J+1)\times(J+1)}$$

elements. The result of the Lemma 5.2.2 implies that the eigenvalues for A can be given as $\gamma^2 + 1 + a^2 - 2a\cos\left(\frac{(k-1)\pi}{J+1}\right)$ where $k = 1, 2, \dots, J+1$, since the largest eigenvalue of A is bounded, the bilinear form $\mathsf{a}(\cdot,\cdot)$ is continuous.
To show coercivity of the bilinear form $\mathsf{a}(\cdot,\cdot)$ we consider the previous rearranged form

$$\begin{aligned}
\mathsf{a}(u,u) &= u^T \mathsf{A} u + (\gamma_0^2 - \gamma^2 - 1 + a)u_0^2 + (a - a^2)u_J^2 \\
&\geq (\gamma^2 + 1 + a^2 - 2a)\|u\|^2 + (\gamma_0^2 - \gamma^2 - 1 + a)u_0^2 + (a - a^2)u_J^2 \\
&= (\gamma_0^2 + a^2 - a)u_0^2 + (\gamma^2 + 1 + a^2 - 2a)\sum_{j=1}^{J-1} u_j^2 + (\gamma^2 + 1 - a)u_J^2. \qquad (5.2.10)
\end{aligned}$$

Setting $c = \min\{(a^2 + \gamma_0^2 - a), (1 + a^2 - 2a + \gamma^2), (\gamma^2 + 1 - a)\}$ gives

$$\mathsf{a}(u,u) \geq c\|u\|^2 \qquad (5.2.11)$$

where $\|\cdot\|$ denotes the Euclidean norm in the space $\mathbb{R}^{J+1}$. For the constant c to be positive we require $\gamma_0^2 > a(1-a)$ and $\gamma^2 > a - 1$. Note that $c > 0$ if $a > 1$ and $\gamma^2 > a - 1$ since $\gamma_0^2 > 0$ under these assumptions. □

Lemma 5.2.2. *Let $\mathsf{B} \in \mathbb{R}^{J\times J}$ be a balanced tridiagonal matrix of the form*

$$\mathsf{B} := \begin{pmatrix}
b-c & -c & 0 & 0 & \dots & 0 & 0 \\
-c & b & -c & 0 & \dots & 0 & 0 \\
0 & -c & b & -c & \dots & 0 & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & 0 & 0 & 0 & \dots & -c & b-c
\end{pmatrix}_{J\times J}$$

then the eigenvalues $\rho^{(k)}$ of the matrix B can be given as $\rho^{(k)} = b - 2c\cos\left(\frac{(k-1)\pi}{J}\right)$ where $k = 1, 2, \dots, J$.

Proof. Lemma 5.2.2 is closely based on Theorem 2 in [84], and follows a similar proof structure. The structure of the matrix allows us to write out the eigenvalue problem

$$\mathsf{B}\omega = \rho\omega$$

equivalently as a set of difference equations (5.2.12). Let $\omega^{(k)} = \left(\omega_1^{(k)}, \omega_2^{(k)}, \dots, \omega_j^{(k)}, \dots, \omega_J^{(k)}\right) \in \mathbb{R}^J$ be the eigenvector corresponding to the k'th eigenvalue of the matrix B, then $\omega^{(k)}$ satisfies the following set of difference equations

$$-c\omega_{j-1}^{(k)} + b\omega_j^{(k)} - c\omega_{j+1}^{(k)} = \rho^{(k)}\omega_j^{(k)} \qquad (5.2.12)$$

where $j = 1, 2, \dots, J$ and boundary conditions $\omega_0^{(k)} = \omega_1^{(k)}$ and $\omega_{J+1}^{(k)} = \omega_J^{(k)}$. We assume the form $\omega_j^{(k)} = \cos\left(\frac{(k-1)(2j-1)\pi}{2J}\right)$ which satisfies the boundary conditions for the difference equations.

Substituting the form of $\omega_j^{(k)}$ in to the equation (5.2.12) and solving for $\rho^{(k)}$ gives

$$\begin{aligned}\rho^{(k)}\cos\Big(\frac{(k-1)(2j-1)\pi}{2J}\Big) &= -c\bigg(\cos\Big(\frac{(k-1)(2j-3)\pi}{2J}\Big)+\cos\Big(\frac{(k-1)(2j+1)\pi}{2J}\Big)\bigg)\\ &+ b\cos\Big(\frac{(k-1)(2j-1)\pi}{2J}\Big)\\ &= -2c\cos\Big(\frac{(k-1)(2j-1)\pi}{2J}\Big)\cos\Big(\frac{(k-1)\pi}{J}\Big)+b\cos\Big(\frac{(k-1)(2j-1)\pi}{2J}\Big).\end{aligned} \tag{5.2.13}$$

Cancelling the $\cos\Big(\frac{(k-1)(2j-1)\pi}{2J}\Big)$ from both sides gives us the required form $\rho^{(k)} = b - 2c\cos\big(\frac{(k-1)\pi}{J}\big)$ for the eigenvalues of the matrix B. □

Properties 5.2.3. *The linear functional* $\mathsf{L}(\cdot):\mathbb{R}^{J+1}\to\mathbb{R}$ *as defined by* (5.2.8) *is continuous.*

Proof. Consider the linear functional $\mathsf{L}(\cdot):\mathbb{R}^{J+1}\to\mathbb{R}$ as

$$\mathsf{L}(u) = \gamma_0^2 m_0 u_0 + \gamma^2\sum_{j=1}^{J}(v_j+\nu_j)u_j \tag{5.2.14}$$

define $r\in\mathbb{R}^{J+1}$ as $r=\big(\gamma_0^2 m_0, \gamma^2(v_1+\nu_1),\ldots,\gamma^2(v_J+\nu_J)\big)$ then we can rewrite $\mathsf{L}(u)$ as

$$\mathsf{L}(u) = \langle r, u\rangle, \tag{5.2.15}$$

where $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product in the space $\mathbb{R}^{J+1}$.The linear functional $\mathsf{L}(\cdot)$ satisfies

$$|\mathsf{L}(u)| \le \|r\|\|u\| \tag{5.2.16}$$

so if $\|r\|<\infty$ the linear functional $\mathsf{L}(\cdot)$ is continuous. □

Now we can state and prove the main result of this section that there exist a unique minimizer of the variational form 5.2.5 which by definition is also the weak constraint 4DVAR filtering solution of the original problem.

Theorem 5.2.4. *Let* $\mathsf{a}(\cdot,\cdot):\mathbb{R}^{J+1}\times\mathbb{R}^{J+1}\to\mathbb{R}$ *be a bilinear form defined by* (5.2.7) *and* $\mathsf{L}(\cdot):\mathbb{R}^{J+1}\to\mathbb{R}$ *be a linear functional as defined by* (5.2.8)*. Let* c *and* r *be defined as in Properties 5.2.1 and 5.2.3. Then there exists a unique* $\hat{u}\in\mathbb{R}^{J+1}$ *such that*

$$\mathsf{a}(\hat{u},\lambda) = \mathsf{L}(\lambda), \quad \forall\lambda\in\mathbb{R}^{J+1},$$

with stability estimate $\|\hat{u}\|\le\frac{\|r\|}{c}$ *and it minimizes the variational functional given by* (5.2.5) *as*

$$\hat{u} = \underset{u\in\mathbb{R}^{J+1}}{\operatorname{argmin}}\,\mathsf{I}^u\big(\{u_j\}_{j=0}^J\big).$$

Proof. Application of Lax-Milgram theorem 5.1.1. □

Now before we analyze the implications of the derived upper bound on the norm of the weak constraint 4DVAR solution we look at the error between the true dynamics v and the weak constraint 4DVAR solution $\hat{u}$. Consider the error $\delta := u - v$ where v is the truth and $u \in \mathbb{R}^{J+1}$. Substituting $u_j = v_j + \delta_j$ in the variational form 5.2.5 yields

$$\begin{aligned}
\mathsf{I}^u\big(\{v_j+\delta_j\}_{j=0}^J\big) &= \frac{\gamma_0^2}{2}(v_0+\delta_0-m_0)^2 + \frac{\gamma^2}{2}\sum_{j=1}^J(v_j+\delta_j-v_j-\nu_j)^2 + \frac{1}{2}\sum_{j=1}^J\big(v_j+\delta_j-a(v_{j-1}+\delta_{j-1})\big)^2 \\
&= \frac{\gamma_0^2}{2}(v_0+\delta_0-m_0)^2 + \frac{\gamma^2}{2}\sum_{j=1}^J(\delta_j-\nu_j)^2 + \frac{1}{2}\sum_{j=1}^J(v_j-av_{j-1}+\delta_j-a\delta_{j-1})^2 \\
&= \frac{\gamma_0^2}{2}(v_0+\delta_0-m_0)^2 + \frac{\gamma^2}{2}\sum_{j=1}^J(\delta_j-\nu_j)^2 + \frac{1}{2}\sum_{j=1}^J(\delta_j-a\delta_{j-1})^2,
\end{aligned}$$

where we apply the equation (5.2.1) to the last term. Owing to the fact that the true solution v is fixed, we can reformulate the variational form 5.2.5 in terms of the error variable, $\mathsf{I}^\delta(\cdot) : \mathbb{R}^{J+1} \to \mathbb{R}$ as

$$\begin{aligned}
\mathsf{I}^\delta\big(\{\delta_j\}_{j=0}^J\big) &= \frac{\sigma^2}{2\sigma_0^2}(v_0-m_0+\delta_0)^2 + \frac{\sigma^2}{2\epsilon^2}\sum_{j=1}^J(\delta_j-\nu_j)^2 + \frac{1}{2}\sum_{j=1}^J(\delta_j-a\delta_{j-1})^2 \qquad (5.2.17)\\
&= \frac{1}{2}\mathsf{a}(\delta,\delta) - \mathsf{L}'(\delta).
\end{aligned}$$

where the bilinear form $\mathsf{a}(\cdot,\cdot)$ is as defined in (5.2.7) and the linear form $\mathsf{L}'(\cdot) : \mathbb{R}^{J+1} \to \mathbb{R}$ is defined as following

$$\mathsf{L}'(\delta) = \gamma_0^2(m_0-v_0)\delta_0 + \gamma^2\sum_{j=1}^J\nu_j\delta_j. \qquad (5.2.18)$$

The Linear functional $\mathsf{L}'(\cdot)$ can be shown to be continuous as was done in the proof of the Property 5.2.3. Define $r' \in \mathbb{R}^{J+1}$ as $r' = \big(\gamma_0^2(m_0-v_0), \gamma^2\nu_1, \ldots, \gamma^2\nu_J\big)$ then we can rewrite $\mathsf{L}'(u)$ as

$$\mathsf{L}'(u) = \langle r', u\rangle, \qquad (5.2.19)$$

where $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product in the space $\mathbb{R}^{J+1}$. The linear functional $\mathsf{L}'(\cdot)$ satisfies

$$|\mathsf{L}'(\delta)| \le \|r'\|\|\delta\| \qquad (5.2.20)$$

so again since $\|r'\| < \infty$ the linear functional $\mathsf{L}'(\cdot)$ is continuous. Since the variational function $\mathsf{I}^\delta(\cdot)$ can be written in terms of the bilinear form $\mathsf{a}(\cdot,\cdot)$ and the linear functional $\mathsf{L}'(\cdot)$ we can again apply the Lax-Milgram result to it.

Theorem 5.2.5. *Let* $\mathsf{a}(\cdot,\cdot) : \mathbb{R}^{J+1}\times\mathbb{R}^{J+1} \to \mathbb{R}$ *be a bilinear form defined by* (5.2.7) *and* $\mathsf{L}'(\cdot)$:

$\mathbb{R}^{J+1} \to \mathbb{R}$ *be a linear functional as defined by* (5.2.18). *Then there exists a unique* $\hat{\delta} \in \mathbb{R}^{J+1}$ *such that*

$$\mathsf{a}(\hat{\delta}, \lambda) = \mathsf{L}'(\lambda), \quad \forall \lambda \in \mathbb{R}^{J+1},$$

with stability estimate $\|\hat{\delta}\| \leq \frac{\|r'\|}{c}$ *and it minimizes the variational functional given by* (5.2.17) *as*

$$\hat{\delta} = \underset{\delta \in \mathbb{R}^{J+1}}{\operatorname{argmin}} \, \mathsf{I}^{\delta}\big(\{\delta_j\}_{j=0}^{J}\big).$$

Proof. Application of Lax-Milgram theorem.(Proposition 5.1.1) □

The following Corollary establishes the linear relation between the minimizers $\hat{u}$ and $\hat{\delta}$.

Corollary 5.2.6. *Let* $\hat{u} = \operatorname{argmin}_{u \in \mathbb{R}^{J+1}} \mathsf{I}^{u}\big(\{u_j\}_{j=0}^{J}\big)$ *and* $\hat{\delta} = \operatorname{argmin}_{\delta \in \mathbb{R}^{J+1}} \mathsf{I}^{\delta}\big(\{\delta_j\}_{j=0}^{J}\big)$ *then* $\hat{\delta} = \hat{u} - v$ *where* v *is the true underlying solution.*

Proof. Assume $\hat{u} - v = \tilde{\delta}$ and $\tilde{\delta} \neq \hat{\delta}$. Let $\lambda \in \mathbb{R}^{J+1}$ be a vector, then from Theorem 5.2.3 we know that $\hat{u}$ satisfies the equation

$$\mathsf{a}(\hat{u}, \lambda) \;\; = \;\; \mathsf{L}(\lambda).$$

On substituting $\hat{u} = \tilde{\delta} + v$, we get

$$\begin{aligned}
\mathsf{a}(\tilde{\delta}, \lambda) &= \mathsf{L}(\lambda) - \mathsf{a}(v, \lambda) \\
&= \gamma_0^2 m_0 \lambda_0 + \gamma^2 \sum_{j=1}^{J} (v_j + \nu_j)\lambda_j - \gamma_0^2 \lambda_0 v_0 - \gamma^2 \sum_{j=1}^{J} \lambda_j v_j \\
&= \gamma_0^2 (m_0 - v_0)\lambda_0 + \gamma^2 \sum_{j=1}^{J} \nu_j \lambda_j, \\
&= \mathsf{L}'(\lambda). \qquad (5.2.21)
\end{aligned}$$

Since λ was chosen arbitrarily it shows that $\tilde{\delta}$ satisfies the equation

$$\mathsf{a}(\tilde{\delta}, \lambda) = \mathsf{L}'(\lambda), \quad \forall \lambda \in \mathbb{R}^{J+1},$$

which contradicts the uniqueness of $\hat{\delta}$. Hence $\hat{u} - v = \hat{\delta}$.

□

Remark 5.2.7. *From the stability estimates of the Theorems 5.2.4 and 5.2.5 we get* $\|\hat{u}\|^2 \leq \frac{\|r\|^2}{c^2}$

and $\|\hat{\delta}\|^2 \leq \frac{\|r'\|^2}{c^2}$. *Taking expectation and using the fact that* $\mathbb{E}[\nu_j^2] = \epsilon^2$ *gives*

$$\mathbb{E}[\|\hat{u}\|^2] \leq \frac{\mathbb{E}[\|r\|^2]}{c^2} = \frac{\gamma_0^4 m_0^2 + J\gamma^4\epsilon^2 + \gamma^4 \sum_{j=1}^{J} v_j^2}{c^2} \tag{5.2.22}$$

$$\mathbb{E}[\|\hat{\delta}\|^2] \leq \frac{\mathbb{E}[\|r'\|^2]}{c^2} = \frac{\gamma_0^4 (v_0 - m_0)^2 + J\gamma^4\epsilon^2}{c^2}. \tag{5.2.23}$$

- *Under the assumptions i)* $a > 1$, *ii)* $\gamma^2 > a - 1$ *and iii)* $\epsilon^2 = \beta\sigma_0^2$, *the expression for the error bound can be rewritten as*

$$\mathbb{E}[\|\hat{\delta}\|^2] \leq \frac{\beta^2\gamma^4(v_0 - m_0)^2 + J\gamma^4\epsilon^2}{\left(\gamma^2 + 1 - a\right)^2}, \quad \textit{when } \beta > 1, \tag{5.2.24}$$

$$\mathbb{E}[\|\hat{\delta}\|^2] \leq \frac{\beta^2\gamma^4(v_0 - m_0)^2 + J\gamma^4\epsilon^2}{\left(\beta\gamma^2 + a^2 - a\right)^2}, \quad \textit{when } \beta < 1. \tag{5.2.25}$$

- *For the parameter* $\beta > 1$ *we get the bounds*

$$\mathbb{E}[\|\hat{\delta}\|^2] \leq \beta^2(v_0 - m_0)^2 + J\epsilon^2, \quad \textit{when } \sigma^2 \to \infty, \tag{5.2.26}$$

$$\mathbb{E}[\|\hat{\delta}\|^2] \leq \beta^2(v_0 - m_0)^2, \quad \textit{when } \epsilon^2 \to 0. \tag{5.2.27}$$

- *For the parameter* $\beta < 1$ *we get the bounds*

$$\mathbb{E}[\|\hat{\delta}\|^2] \leq (v_0 - m_0)^2 + \frac{J\epsilon^2}{\beta^2}, \quad \textit{when } \sigma^2 \to \infty, \tag{5.2.28}$$

$$\mathbb{E}[\|\hat{\delta}\|^2] \leq (v_0 - m_0)^2, \quad \textit{when } \epsilon^2 \to 0. \tag{5.2.29}$$

These assumptions in conjunction with the stability estimate show us that when the model error is large and observational noise are small the mean squared error for the estimate from the Weak 4DVAR algorithm only depends on the error in the estimation of the initial condition.

5.3 3DVAR Constraint 4DVAR with Model Error

In this section we again consider the underlying system v and observation y as given by the equation (5.2.2). However in place of the model dynamics (5.2.3), we choose the following linear one dimensional system with the 3DVAR innovation term

$$u_j = \Psi(u_{j-1}) + g(y_j - au_{j-1}) := au_{j-1} + \xi_j + g(y_j - au_{j-1}) \tag{5.3.1}$$

with the initial condition $u_0 \sim N(m_0, \sigma_0^2)$, the model error $\xi_j \sim N(0, \sigma^2)$ and where $g \in \mathbb{R}$ is the Kalman gain coefficient. For 3DVAR sequential filtering scheme we choose $g = \frac{\sigma_0^2}{\sigma_0^2+\epsilon^2}$. Now we define the weak constraint 4DVAR minimization functional $\tilde{\mathsf{I}}(\cdot,\cdot) : \mathbb{R}^{J+1} \times \mathbb{R}^J \to \mathbb{R}$ as before:

$$\tilde{\mathsf{I}}\big(u, \xi\big) = \frac{1}{2\sigma_0^2}(u_0 - m_0)^2 + \frac{1}{2\epsilon^2}\sum_{j=1}^{J}(u_j - y_j)^2 + \frac{1}{2\sigma^2}\sum_{j=1}^{J}\xi_j^2. \tag{5.3.2}$$

The minimization of the cost function again is constrained by the dynamics (5.3.1) of the state variable u, so we reformulate the minimization functional as $\tilde{\mathsf{I}}^u(\cdot) : \mathbb{R}^{J+1} \to \mathbb{R}$, in terms of the state variable as

$$\tilde{\mathsf{I}}^u\big(\{u_j\}_{j=0}^J\big) = \frac{1}{2\sigma_0^2}(u_0-m_0)^2+\frac{1}{2\epsilon^2}\sum_{j=1}^{J}(u_j-y_j)^2+\frac{1}{2\sigma^2}\sum_{j=1}^{J}(u_j-a(1-g)u_{j-1}-agv_{j-1}-g\nu_j)^2. \tag{5.3.3}$$

Again the state vector u which minimizes the variational form (5.3.3) is the 3DVAR modified weak constraint 4DVAR filtering solution for the underlying dynamical system. On simplifying and rewriting the variational form (5.3.3) in terms of the parameters $\gamma_0^2 = \frac{\sigma^2}{\sigma_0^2}$, $\gamma^2 = \frac{\sigma^2}{\epsilon^2}$ we get

$$\begin{aligned}
\tilde{\mathsf{I}}^u\big(\{u_j\}_{j=0}^J\big) &= \frac{\sigma^2}{2\sigma_0^2}(u_0 - m_0)^2 + \frac{\sigma^2}{2\epsilon^2}\sum_{j=1}^{J}(u_j - v_j - \nu_j)^2 + \frac{1}{2}\sum_{j=1}^{J}(u_j - a(1-g)u_{j-1} - agv_j - g\nu_j)^2 \\
&= \frac{\gamma_0^2}{2}(u_0 - m_0)^2 + \frac{\gamma^2}{2}\sum_{j=1}^{J}(u_j - v_j - \nu_j)^2 + \frac{1}{2}\sum_{j=1}^{J}(u_j - a(1-g)u_{j-1} - agv_j - g\nu_j)^2 \\
&= \frac{1}{2}\tilde{\mathsf{a}}(u,u) - \tilde{\mathsf{L}}(u).
\end{aligned} \tag{5.3.4}$$

where the bilinear form $\tilde{\mathsf{a}}(\cdot,\cdot)$ and the linear functional $\tilde{\mathsf{L}}(\cdot)$ are defined as following

$$\tilde{\mathsf{a}}(u,\lambda) = \gamma_0^2 u_0\lambda_0 + \gamma^2\sum_{j=1}^{J}u_j\lambda_j + \sum_{j=1}^{J}(u_j - a(1-g)u_{j-1})(\lambda_j - a(1-g)\lambda_{j-1}) \tag{5.3.5}$$

$$\tilde{\mathsf{L}}(\lambda) = \gamma_0^2 m_0\lambda_0 + \Big(\gamma^2 + g\Big)\sum_{j=1}^{J}(v_j + \nu_j)\lambda_j - \sum_{j=1}^{J}ag(1-g)(v_j + \nu_j)\lambda_{j-1}. \tag{5.3.6}$$

Now we establish the Properties 5.2.1 and 5.2.3 for the bilinear form $\tilde{\mathsf{a}}(\cdot,\cdot)$ and the linear functional $\tilde{\mathsf{L}}(\cdot)$ respectively.

Properties 5.3.1. *The bilinear form $\tilde{\mathsf{a}}(\cdot,\cdot) : \mathbb{R}^{J+1} \times \mathbb{R}^{J+1} \to \mathbb{R}$ as defined by* (5.3.5) *is symmetric, continuous and coercive.*

Proof. The bilinear form $\tilde{\mathsf{a}}(\cdot,\cdot)$ as defined in the equation (5.3.5) is clearly symmetric. To prove

the continuity of the bilinear form, consider the following rearrangement of $\tilde{\mathsf{a}}(\cdot,\cdot)$ as

$$\begin{aligned}
\tilde{\mathsf{a}}(u,\lambda) &= \gamma_0^2 u_0\lambda_0 + \gamma^2\sum_{j=1}^{J} u_j\lambda_j + \sum_{j=1}^{J}(u_j - a(1-g)u_{j-1})(\lambda_j - a(1-g)\lambda_{j-1}) \\
&= \gamma_0^2 u_0\lambda_0 + \gamma^2\sum_{j=1}^{J} u_j\lambda_j + \sum_{j=1}^{J}\big(u_j\lambda_j - a(1-g)u_{j-1}\lambda_j - a(1-g)u_j\lambda_{j-1} + a(1-g)^2u_{j-1}\lambda_{j-1}\big) \\
&= (\gamma_0^2+1)u_0\lambda_0 + (\gamma^2+1+a^2(1-g)^2)\sum_{j=1}^{J-1} u_j\lambda_j + (\gamma^2+1)u_J\lambda_J - a(1-g)\sum_{j=1}^{J}\big(u_j\lambda_{j-1} + u_{j-1}\lambda_j\big) \\
&= u^T\tilde{\mathsf{A}}\lambda. \qquad (5.3.7)
\end{aligned}$$

Again the eigenvalues of this matrix can be given as $\Theta + 2a(1-g)\cos(\frac{k\pi}{J+2})$ where $k = 1,2,\ldots,J+1$ and $\Theta := \max\{(\gamma_0^2 + a^2(1-g)^2), (\gamma^2 + a^2(1-g)^2 + 1)\}$, since the largest eigenvalue of $\tilde{\mathsf{A}}$ is bounded, the bilinear form $\tilde{\mathsf{a}}(\cdot,\cdot)$ is continuous.
To show coercivity of the bilinear form $\tilde{\mathsf{a}}(\cdot,\cdot)$ we consider the following rearranged form as following

$$\begin{aligned}
\tilde{\mathsf{a}}(u,u) &\geq \big(a^2(1-g)^2 + \gamma_0^2 - a(1-g)\big)u_0^2 \\
&+ \big((1-a(1-g))^2 + \gamma^2\big)\sum_{j=1}^{J-1} u_j^2 + (\gamma^2 + 1 - a(1-g))u_J^2. \qquad (5.3.8)
\end{aligned}$$

Setting $\tilde{c} = \min\{\big(a^2(1-g)^2 + \gamma_0^2 - a(1-g)\big), \big((1-a(1-g))^2 + \gamma^2\big), (\gamma^2 + 1 - a(1-g))\}$ gives

$$\tilde{\mathsf{a}}(u,u) \geq \tilde{c}\|u\|^2. \qquad (5.3.9)$$

where $\|\cdot\|$ denotes the Euclidean norm in the space $\mathbb{R}^{J+1}$. For the constant $\tilde{c}$ to be positive we require $\gamma_0^2 > a(1-g)\big(1-a(1-g)\big)$ and $\gamma^2 > 1 - a(1-g)$. Note that $\tilde{c} > 0$ if $a(1-g) > 1$ since $\gamma_0^2, \gamma^2 > 0$ by definition. □

Properties 5.3.2. *The linear functional $\tilde{\mathsf{L}}(\cdot) : \mathbb{R}^{J+1} \to \mathbb{R}$ as defined by* (5.3.6) *is continuous.*

Proof. Consider the linear functional $\tilde{\mathsf{L}}(\cdot) : \mathbb{R}^{J+1} \to \mathbb{R}$ as

$$\tilde{\mathsf{L}}(u) = \gamma_0^2 u_0 m_0 + (\gamma^2 + g)\sum_{j=1}^{J} u_j(v_j + \nu_j) - ag(1-g)\sum_{j=1}^{J}(v_j + \nu_j)u_{j-1} \qquad (5.3.10)$$

define $\tilde{r} \in \mathbb{R}^{J+1}$ as

$$\begin{aligned}
\tilde{r} = \Big(&\gamma_0^2 m_0 - ag(1-g)(v_1+\nu_1), (\gamma^2+g)(v_1+\nu_1) - ag(1-g)(v_2+\nu_2), \ldots \\
&, \underbrace{(\gamma^2+g)(v_j+\nu_j) - ag(1-g)(v_{j+1}+\nu_{j+1})}_{j+1\text{th term}}, \ldots, (\gamma^2+g)(v_J+\nu_J)\Big) \qquad (5.3.11)
\end{aligned}$$

then we can rewrite $\tilde{\mathsf{L}}(u)$ as

$$\tilde{\mathsf{L}}(u) = \langle \tilde{r}, u \rangle, \tag{5.3.12}$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product in the space $\mathbb{R}^{J+1}$.The linear functional $\tilde{\mathsf{L}}(\cdot)$ satisfies

$$|\tilde{\mathsf{L}}(u)| \leq \|\tilde{r}\| \|u\| \tag{5.3.13}$$

so if $\|\tilde{r}\| < \infty$ the linear functional $\tilde{\mathsf{L}}(\cdot)$ is continuous. □

Now we can state and prove a result similar to the Theorem 5.2.4 that there exist a unique minimizer of the variational form (5.3.3) which by definition is also the 3DVAR modified weak constraint 4DVAR filtering solution of the original problem.

Theorem 5.3.3. *Let $\tilde{\mathsf{a}}(\cdot, \cdot) : \mathbb{R}^{J+1} \times \mathbb{R}^{J+1} \to \mathbb{R}$ be a bilinear form defined by* (5.3.5) *and $\tilde{\mathsf{L}}(\cdot) : \mathbb{R}^{J+1} \to \mathbb{R}$ be a linear functional as defined by* (5.3.6)*. Then there exists a unique $\hat{u} \in \mathbb{R}^{J+1}$ such that*

$$\tilde{\mathsf{a}}(\hat{u}, \lambda) = \tilde{\mathsf{L}}(\lambda), \quad \forall \lambda \in \mathbb{R}^{J+1},$$

with stability estimate $\|\hat{u}\| \leq \frac{\|\tilde{r}\|}{\tilde{c}}$ and it minimizes the variational functional given by (5.3.3) *as*

$$\hat{u} = \underset{u \in \mathbb{R}^{J+1}}{\operatorname{argmin}} \tilde{\mathsf{I}}^u\big(\{u_j\}_{j=0}^J\big).$$

Proof. Application of Lax-Milgram theorem. □

To obtain the error bound on the solution of modified weak constraint 4DVAR minimization we consider the error variable $\delta := u - v$ where v is the truth and $u \in \mathbb{R}^{J+1}$. A reformulation of the variational form (5.3.3) in terms of the error variable, $\tilde{\mathsf{I}}^\delta(\cdot) : \mathbb{R}^{J+1} \to \mathbb{R}$ can be given as

$$\begin{aligned}
\mathsf{I}^\delta\big(\{\delta_j\}_{j=0}^J\big) &= \frac{\sigma^2}{2\sigma_0^2}(v_0 - m_0 + \delta_0)^2 + \frac{\sigma^2}{2\epsilon^2}\sum_{j=1}^J(\delta_j - \nu_j)^2 + \frac{1}{2}\sum_{j=1}^J(\delta_j - a(1-g)\delta_{j-1} - g\nu_j)^2 \\
&= \frac{\gamma_0^2}{2}(v_0 - m_0 + \delta_0)^2 + \frac{\gamma^2}{2}\sum_{j=1}^J(\delta_j - \nu_j)^2 + \frac{1}{2}\sum_{j=1}^J(\delta_j - a(1-g)\delta_{j-1} - g\nu_j)^2 \\
&= \frac{1}{2}\tilde{\mathsf{a}}(\delta, \delta) - \tilde{\mathsf{L}}'(\delta).
\end{aligned} \tag{5.3.14}$$

where the bilinear form $\tilde{\mathsf{a}}(\cdot, \cdot)$ is as defined in (5.3.5) and the linear form $\tilde{\mathsf{L}}'(\cdot) : \mathbb{R}^{J+1} \to \mathbb{R}$ is defined as following

$$\tilde{\mathsf{L}}'(\delta) = -\gamma_0^2(v_0 - m_0)\delta_0 + (\gamma^2 + g)\sum_{j=1}^J \nu_j\delta_j - ag(1-g)\sum_{j=1}^J \nu_j\delta_{j-1}. \tag{5.3.15}$$

The Linear functional $\tilde{\mathsf{L}}'(\cdot)$ can be shown to be continuous as was done in the proof of the Property

5.3.2. Define $r' \in \mathbb{R}^{J+1}$ as

$$\begin{aligned}\tilde{r}' &= \Big(-\gamma_0^2(v_0 - m_0) - ag(1-g)\nu_1, (\gamma^2+g)\nu_1 - ag(1-g)\nu_2, \\ &\qquad \ldots, \underbrace{(\gamma^2+g)\nu_j - ag(1-g)\nu_{j+1}}_{j+1\text{th term}}, \ldots, (\gamma^2+g)\nu_J \Big) \end{aligned} \tag{5.3.16}$$

then we can rewrite $\tilde{\mathsf{L}}'(u)$ as

$$\tilde{\mathsf{L}}'(u) = \langle \tilde{r}', u \rangle, \tag{5.3.17}$$

and the linear functional $\tilde{\mathsf{L}}'(\cdot)$ satisfies

$$|\tilde{\mathsf{L}}'(\delta)| \leq \|\tilde{r}'\| \|\delta\|. \tag{5.3.18}$$

So again if $\|\tilde{r}'\| < \infty$ the linear functional $\tilde{\mathsf{L}}'(\cdot)$ is continuous. Since the variational function $\tilde{\mathsf{I}}^{\delta}(\cdot)$ can be written in terms of the bilinear form $\tilde{\mathsf{a}}(\cdot,\cdot)$ and the linear functional $\tilde{\mathsf{L}}'(\cdot)$ we can again apply the Lax-Milgram result to it.

Theorem 5.3.4. *Let $\tilde{\mathsf{a}}(\cdot,\cdot) : \mathbb{R}^{J+1} \times \mathbb{R}^{J+1} \to \mathbb{R}$ be a bilinear form defined by* (5.3.5) *and $\tilde{\mathsf{L}}'(\cdot) : \mathbb{R}^{J+1} \to \mathbb{R}$ be a linear functional as defined by* (5.3.15)*. Then there exists a unique $\hat{\delta} \in \mathbb{R}^{J+1}$ such that*

$$\tilde{\mathsf{a}}(\hat{\delta}, \lambda) = \tilde{\mathsf{L}}'(\lambda), \quad \forall \lambda \in \mathbb{R}^{J+1},$$

with stability estimate $\|\hat{\delta}\| \leq \frac{\|\tilde{r}'\|}{c}$ and it minimizes the variational functional given by (5.3.4) *as*

$$\hat{\delta} = \underset{\delta \in \mathbb{R}^{J+1}}{\operatorname{argmin}} \tilde{\mathsf{I}}^{\delta}\big(\{\delta_j\}_{j=0}^{J}\big).$$

Proof. Application of Lax-Milgram theorem.(Proposition 5.1.1) □

The final step is to establish the linear relation between the minimizers $\hat{u}$ and $\hat{\delta}$.

Corollary 5.3.5. *Let $\hat{u} = \operatorname{argmin}_{u \in \mathbb{R}^{J+1}} \tilde{\mathsf{I}}^{u}\big(\{u_j\}_{j=0}^{J}\big)$ and $\hat{\delta} = \operatorname{argmin}_{\delta \in \mathbb{R}^{J+1}} \tilde{\mathsf{I}}^{\delta}\big(\{\delta_j\}_{j=0}^{J}\big)$ then $\hat{\delta} = \hat{u} - v$ where v is the true underlying solution.*

Proof. Assume $\hat{u} - v = \tilde{\delta}$ and $\tilde{\delta} \neq \hat{\delta}$. Let $\lambda \in \mathbb{R}^{J+1}$ be a given vector then from Theorem 5.3.3 we know that $\hat{u}$ satisfies the equation

$$\tilde{\mathsf{a}}(\hat{u}, \lambda) \;=\; \tilde{\mathsf{L}}(\lambda).$$

On substituting $\hat{u} = \tilde{\delta} + v$, we get

$$\begin{aligned}
\tilde{\mathsf{a}}(\tilde{\delta}, \lambda) &= \tilde{\mathsf{L}}(\lambda) - \tilde{\mathsf{a}}(v, \lambda) \\
&= \gamma_0^2\lambda_0 m_0 + (\gamma^2+g)\sum_{j=1}^{J}\lambda_j(v_j+\nu_j) - ag(1-g)\sum_{j=1}^{J}(v_j+\nu_j)\lambda_{j-1} - \gamma_0^2\lambda_0 v_0 - \gamma^2\sum_{j=1}^{J}\lambda_j v_j \\
&\quad -\sum_{j=1}^{J}(\lambda_j - a(1-g)\lambda_{j-1})(v_j - a(1-g)v_{j-1}) \\
&= \gamma_0^2\lambda_0 m_0 + (\gamma^2+g)\sum_{j=1}^{J}\lambda_j(v_j+\nu_j) - ag(1-g)\sum_{j=1}^{J}(v_j+\nu_j)\lambda_{j-1} - \gamma_0^2\lambda_0 v_0 - \gamma^2\sum_{j=1}^{J}\lambda_j v_j \\
&\quad -\sum_{j=1}^{J}(\lambda_j - a(1-g)\lambda_{j-1})g v_j \\
&= \gamma_0^2(m_0 - v_0)\lambda_0 + (\gamma^2+g)\sum_{j=1}^{J}\nu_j\lambda_j - ag(1-g)\sum_{j=1}^{J}\nu_j\lambda_{j-1}, \\
&= \tilde{\mathsf{L}}'(\lambda). \qquad (5.3.19)
\end{aligned}$$

Since λ was chosen arbitrarily it shows that $\tilde{\delta}$ satisfies the equation

$$\tilde{\mathsf{a}}(\tilde{\delta}, \lambda) = \tilde{\mathsf{L}}'(\lambda), \quad \forall \lambda \in \mathbb{R}^{J+1},$$

which contradicts the uniqueness of $\hat{\delta}$. Hence $\hat{u} - v = \hat{\delta}$.

□

Remark 5.3.6. *From the stability estimates of the Theorems 5.3.3 and 5.3.4 we get* $\|\hat{u}\|^2 \leq \frac{\|\tilde{r}\|^2}{\tilde{c}^2}$ *and* $\|\hat{\delta}\|^2 \leq \frac{\|\tilde{r}'\|^2}{\tilde{c}^2}$. *Taking expectation and using the notation* $\alpha := ag(1-g)$ *and* $\mathbb{E}[\nu_j^2] = \epsilon^2$ *gives*

$$\begin{aligned}
\mathbb{E}[\|\hat{u}\|^2] \leq \frac{\mathbb{E}[\|\tilde{r}\|^2]}{c^2} &= \frac{\left(\gamma_0^2 m_0 - \alpha v_1\right)^2 + J\left((\gamma^2+g)^2 + \alpha^2\right)\epsilon^2 + \sum_{j=1}^{J-1}\left((\gamma^2+g)v_j - \alpha v_{j+1}\right)^2 + \left(\gamma^2+g\right)^2 v_J^2}{\tilde{c}^2} \\
\mathbb{E}[\|\hat{\delta}\|^2] \leq \frac{\mathbb{E}[\|\tilde{r}'\|^2]}{\tilde{c}^2} &= \frac{\gamma_0^4(v_0 - m_0)^2 + J((\gamma^2+g)^2 + \alpha^2)\epsilon^2}{\tilde{c}^2}.
\end{aligned}$$

- *Under the assumptions i)* $\tilde{a} > 1$, *ii)* $\gamma^2 > \tilde{a} - 1$ *and iii)* $\epsilon^2 = \beta\sigma_0^2$ *the expression for error can be rewritten as*

$$\mathbb{E}[\|\hat{\delta}\|^2] \leq \frac{\beta^2\gamma^4(v_0 - m_0)^2 + J\gamma^4\epsilon^2}{\left(\gamma^2 + 1 - a\right)^2}, \quad \textit{when } \beta > 1, \qquad (5.3.20)$$

$$\mathbb{E}[\|\hat{\delta}\|^2] \leq \frac{\beta^2\gamma^4(v_0 - m_0)^2 + J\gamma^4\epsilon^2}{\left(\beta\gamma^2 + a^2 - a\right)^2}, \quad \textit{when } \beta < 1. \qquad (5.3.21)$$

- *When the model errors are large i.e.* $\sigma^2 \to \infty$ *we get*

$$\mathbb{E}[\|\hat{\delta}\|^2] \quad \leq \quad \beta^2(v_0 - m_0)^2 + J\epsilon^2, \quad \textit{when } \beta > 1, \tag{5.3.22}$$

  i.e. the model errors effectively do not contribute to the upper bound on the error.

- *If we make assumptions of small observation noise* $\epsilon^2 \to 0$ *we get the bound*

$$\mathbb{E}[\|\hat{\delta}\|^2] \quad \leq \quad \beta^2(v_0 - m_0)^2, \quad \textit{when } \beta > 1. \tag{5.3.23}$$

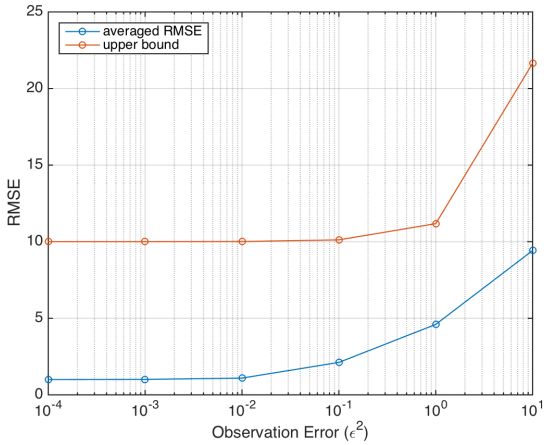
5.4 Numerical Results: Model Noise

In this section we numerically demonstrate the theoretical results presented in previous section. In the first part of this section we consider a linear model and then extend the results to Lorenz'63 model.

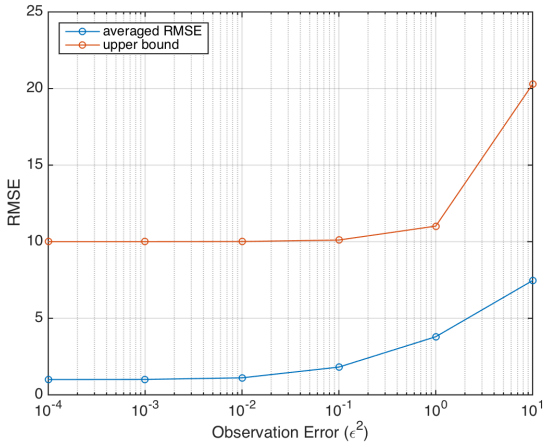
5.4.1 Linear Model

The weak constraint experiments for the linear systems follow a similar set up as for the strong constraint experiments, in that is the growth coefficient is chosen to be $a = 1.2$ in equation (5.2.1), the observation errors are distributed as $N(0, \epsilon^2)$. The model errors are generated randomly from a Gaussian distribution $N(0, \sigma^2)$, with a specific, defined variance, and zero mean. These are then added to the model equation at each time step. The true initial condition is chosen to be $v_0 = 1$ and the model initial condition is chosen to be $m_0 = 2$ for all the linear experiments. The results are presented for different combinations of model and observation error values.

The first experiment is where the model error variance is fixed to $\sigma^2 = 10$ and the observation errors are varied. The results for this experiment are shown in Figure 5.4.1a for weak 4DVAR scheme and in figure 5.4.1b for 3DVAR constrained weak 4DVAR scheme. We observe that in both the cases the error in the estimate stays below the theoretical upper bound established in sections 5.2 and 5.3. We see that as the observation error becomes small, the theoretical error estimates for both the schemes converge to the same limit given by the expressions given in Remarks 5.2.7 and 5.3.6 for $\epsilon^2 \to 0$. Also notice that RMSE is lower for 3DVAR constraint 4DVAR for the same set of parameters.



(a) Standard Weak constraint 4DVAR

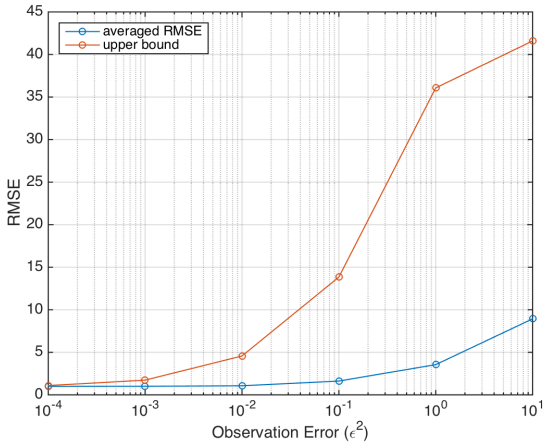


(b) 3DVAR constraint 4DVAR

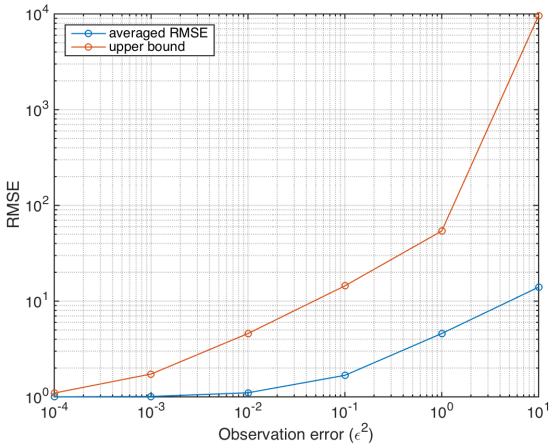
Figure 5.4.1: The parameter values chosen for this experiments are $\beta^2 = 10$, $J = 20$ and $\sigma^2 = 10$.

We next consider the case when the ratio of the variance in the initial guess σ_0^2 and the observation noise ϵ^2, $\beta := \frac{\epsilon^2}{\sigma_0^2}$ is less than one. The results presented show that as the model and observations are accurate becomes the result closes

Again we see that as the observation error becomes small, the theoretical error estimates for both the schemes converge to the same limit given by the expressions given in Remarks 5.2.7 and 5.3.6 for $\epsilon^2 \to 0$. Since the background error is smaller than the observation error we see lower error for accurate observations how ever when the observation error is large the upper bound is large as well. In this case accuracy of 3DVAR scheme is better compared to the case when $\beta^2 = 10$ so we observe larger error in Figure 5.4.2b.



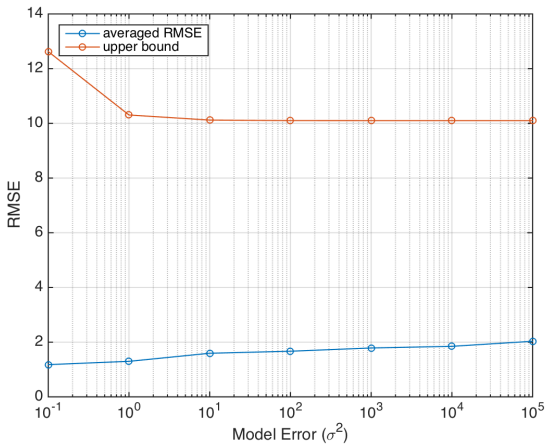
(a) Standard Weak constraint 4DVAR



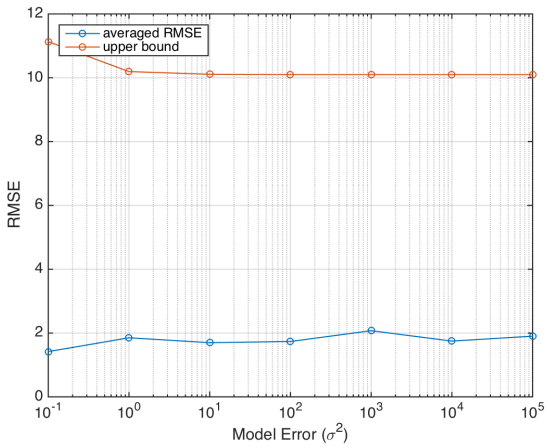
(b) 3DVAR constraint 4DVAR

Figure 5.4.2: The parameter values chosen for this experiments are $\beta^2 = 0.1$, $J = 20$ and $\sigma^2 = 10$. Note the Figure 5.4.2b is log-log plot.

In the next set of experiments we fix the observation error $\epsilon^2 = 0.1$ and vary the level of model error present. The results for this experiment are shown in Figure 5.4.3a for weak 4DVAR scheme and in figure 5.4.3b for 3DVAR constrained 4DVAR scheme. We observe that in both the cases the error in the estimate stays below the theoretical upper bound established in section 5.2 and 5.3. We observe that for small model noise the bound is higher and as the model error grows large the error bound approaches the asymptotic value, as the model noise term in the cost function becomes less significant.



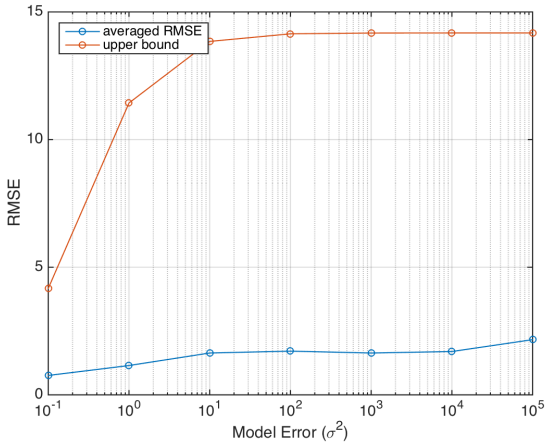
(a) Standard Weak constraint 4DVAR



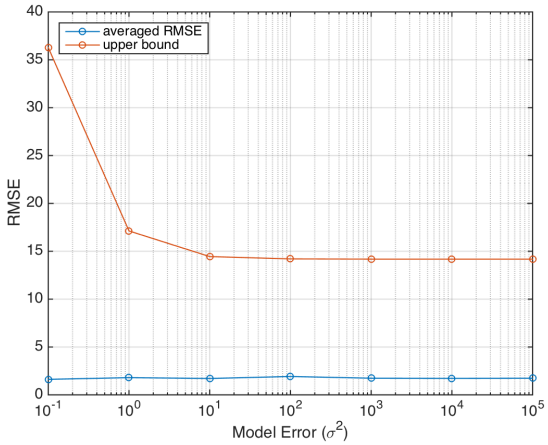
(b) 3DVAR constraint 4DVAR

Figure 5.4.3: The parameter values chosen for this experiments are $\beta^2 = 10$, $J = 20$ and $\epsilon^2 = 0.1$. We see that as the model error becomes large the theoretical error estimates for both the schemes converge to the same limit given by the expressions given in Remarks 5.2.7 and 5.3.6 for $\sigma^2 \to \infty$.

In the final experiment we consider the ratio of the variance in the initial guess σ_0^2 and the observation noise ϵ^2, $\beta := \frac{\epsilon^2}{\sigma_0^2}$ to be 0.1. We vary the model noise σ^2 and the results are presented in the Figure 5.4.4. The asymptotic error upper bounds approach the same level as seen in Figures 5.4.4a and 5.4.4b.



(a) Standard Weak constraint 4DVAR



(b) 3DVAR constraint 4DVAR

Figure 5.4.4: The parameter values chosen for this experiments are $\beta^2 = 0.1$, $J = 20$ and $\epsilon^2 = 0.1$. Similarly in this case also see that as the model error becomes large the theoretical error estimates for both the schemes converge to the same limit given by the expressions given in Remarks 5.2.7 and 5.3.6 for $\sigma^2 \to \infty$.

5.4.2 Nonlinear Models

In this section we focus on nonlinear chaotic models. We compare the errors in trajectory estimates provided by weak constraint 4DVAR and 3DVAR constraint weak 4DVAR. The application of 3DVAR constraint 4DVAR algorithm behaves differently depending upon the accuracy of the 3DVAR filter. The averaged root mean squared error (RMSE) is used to evaluate the accuracy of the 3DVAR scheme. For N dimensional system with true state $v = (v^{(1)}, \ldots, v^{(N)})$, J observations steps and the filtering estimate $u = (u^{(1)}, \ldots, u^{(N)})$, RMSE for one instance of error process can be expressed as following

$$RMSE = \frac{1}{J}\sum_{j=1}^{J}\sqrt{\frac{1}{N}\sum_{n=1}^{N}(v_j^{(n)} - u_j^{(n)})^2}. \tag{5.4.1}$$

The RMSE shown in the results is averaged over 100 instances of error process. To minimize the cost functions, the MATLAB routine FMINSEARCH, which uses a Nelder-Mead simplex direct search method to find the minimum, is used.

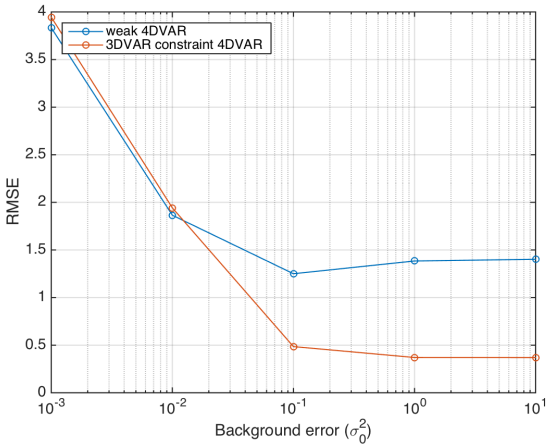
Lorenz'63 Model

We consider the Lorenz'63 equations [53], as described in section 4.5.2. The true trajectory is calculated by using the fourth-order Runge-Kutta method. The model equation are integrated using modified Euler scheme with step size $\triangle t = 0.01$. We define the background matrix as $C_0 := \sigma^2 \mathbb{I}_{3\times 3}$. To evaluate the performance of the method, we use the twin experiment technique. The observations are generated at the observation times by adding mean zero Gaussian noise with the covariance matrix $\Gamma := \epsilon^2 \mathbb{I}_{3\times 3}$ to the underlying truth trajectory.

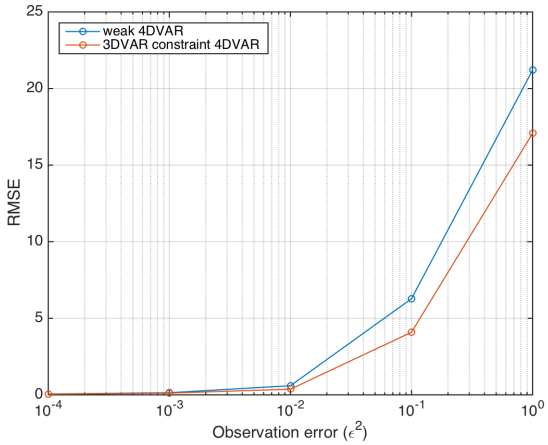
We observe all the components of the three dimensional system in observation intervals of $h = 0.1$, i.e., $t_j = 0.1j$. A total of $J = 50$ assimilation steps are performed. The model error is implemented by adding zero mean Gaussian noise to the model trajectory at each time step. The covariance matrix for the model errors is given as $\Sigma := \sigma^2 \mathbb{I}_{3\times 3}$.

Figure 5.4.5 reports simulation results for assimilating observations over 50 assimilation cycles, using the weak constraint 4DVAR and 3DVAR constraint weak 4DVAR method. Assimilation window length is chosen to be $\triangle t = 0.5$. In Figure 5.4.5a we plot averaged RMSE against the background variance, keeping the observation error ($\epsilon^2 = 10^{-2}$) and model error ($\sigma^2 = 0.1$) fixed. We observe that for smaller values of the background variance both the schemes perform almost the same. However, when the value of σ_0^2 is chosen to be large enough that the 3DVAR filter tracks the underlying system accurately the 3DVAR constraint weak 4DVAR out performs the standard weak constraint 4DVAR.

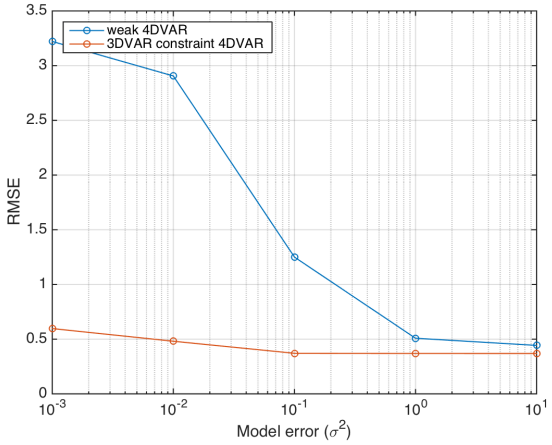
In Figure 5.4.5b we plot averaged RMSE against the observation error ϵ^2, keeping the background variance ($\sigma_0^2 = 10^{-2}$) and model error ($\sigma^2 = 0.1$) fixed. In this case for smaller values of ϵ^2, the observations are accurate and both the schemes perform equally. As the observation noise increases the difference between two schemes becomes more prominent. Although for large noise 3DVAR filter fails to accurately approximate the underlying system hence the RMSE also grows large. In Figure 5.4.5c we plot averaged RMSE against the model error σ^2, keeping the background variance ($\sigma_0^2 = 10^{-1}$) and observation error ($\epsilon^2 = 10^{-2}$) fixed. For chosen values of background variance and observation error the 3DVAR filter accurately tracks the system. We see in the Figure 5.4.5c, in case of standard weak 4DVAR, for smaller model error values the model error term dominates the cost function as, although the added model noise terms are small, incorrect initial condition and coarse discretization keeps the term large. Whereas in the case of 3DVAR constraint weak 4DVAR the model error term is kept small by 3DVAR scheme accurately tracking the trajectory. However, for the large model error values the model error term becomes less significant and the standard weak 4DVAR scheme converges to the similar RMSE values as the 3DVAR constrained weak 4DVAR.



(a) $\sigma^2 = 0.1$ and $\epsilon^2 = 10^{-2}$



(b) $\sigma_0^2 = 0.1$ and $\sigma^2 = 0.1$



(c) $\sigma_0^2 = 0.1$ and $\epsilon^2 = 10^{-2}$

Figure 5.4.5: Comparison of RMSE results of weak 4DVAR and 3DVAR constrained 4DVAR schemes for Lorenz'63 Model.

5.5 Conclusions

In this chapter, the traditional data assimilation method weak constraint 4DVAR is compared against the hybrid method 3DVAR constraint weak 4DVAR. For linear system we observed that analytical upper bound provided in the theory holds for a range of combinations of model error and observation error values. When the data is accurate i.e. the observation error $\epsilon^2 \approx 0$, the upper bounds provided by both standard weak 4DVAR and 3DVAR constraint weak 4DVAR are dominated by the error present in the initial condition. In case of nonlinear model it has been shown that over smaller data assimilation windows, when the error growth is close to linear, 3DVAR constraint weak 4DVAR outperforms standard weak 4DVAR method when the 3DVAR scheme is accurate. As the ratio of observation noise to background increases the sequential filter estimate becomes more reliant on the background model which can lead to inaccurate 3DVAR filter for some

regimes.

However, some of the key questions not attempted in this work are the effect of this scheme in larger assimilation windows [72], its comparison with ensemble based hybrid schemes [25] and formulation of appropriate background covariance matrix [77].

Still, it can be seen that variational methods when combined with 3DVAR sequential filter improve the accuracy of the forecast.

Chapter 6

Conclusion and Future Work

Data assimilation is a method that combines model dynamics, state observations and the error statistics to obtain good estimations of the system state. Approximate Gaussian filters are some of the most common algorithms used for data assimilation problems. In this thesis we investigated the role of ideas from dynamical systems in the rigorous analysis of filtering schemes and, through computational studies shows the gap between theory and practice, demonstrating the need for further theoretical developments.

In the following sections we present a summary of the main results of this work and discuss possible further lines of investigation.

6.1 Conclusions

In Chapter 2 we studied the long-time behaviour of filters for partially observed dissipative dynamical systems and the properties of the 3DVAR algorithm when applied to the partially observed Lorenz '63 model extending the more involved theory developed for the 3DVAR filter applied to the partially observed Navier-Stokes equations in [11, 8].

In Chapter 3 we have highlighted the connection to synchronization in dynamical systems, and shown that this synchronization theory, which applies to noise-free data, is robust to the addition of noise, in both the continuous and discrete time settings. We also extend the accuracy results from Chapter 2 to the Lorenz'96 System with the 3DVAR algorithm. In the context of the Lorenz '96 model we have identified a fixed observation operator, based on observing 2/3 of the components of the signal's vector sufficient to ensure desirable long-time properties of the filter. We also studied adaptive observation operators, targeted to observe the directions of maximal growth within the local linearized dynamics. We demonstrated that with these adaptive observers, considerably fewer observations are required. We also draw comparison between these adaptive observation operators, and the AUS methodology which is also based on the local linearized dynamics, but works by projecting within the model covariance operators of ExKF, whilst the observation operators themselves are fixed; thus the model covariances are adapted. Both adaptive observation operators and the

AUS methodology show the potential for considerable computational savings in filtering, without loss of accuracy. Although, the adaptive observation operator methods may not be implementable in practice on the high dimensional systems arising in, for example, meteorological applications, they provide conceptual insights into the development of improved algorithms.

In Chapter 4 we introduced a 4DVAR assimilation scheme constrained by 3DVAR estimates. We applied the aforementioned scheme to a simple linear state data assimilation problem and established the existence of the bias in the initial condition for both growing and contracting linear systems. We also established the almost sure convergence of the 4DVAR estimate for initial condition using the path integral approach. Furthermore, under the path integral framework, we showed that the asymptotic variance of the estimate of initial condition from 3DVAR constraint 4DVAR formulation agrees with the bias estimate given by the mean square convergence result.

The two main challenges in implementing the 3DVAR constrained 4DVAR scheme are first to find the appropriate Kalman gain factor where 3DVAR accurately tracks the underlying process without over-saturating the underlying signal. The second challenge is the pre-computation of the bias given the model and the filtering parameters especially for more complex non-linear models which remains an open problem.

In Chapter 5 we extended the 3DVAR constraint to weak 4DVAR scheme. For linear system we derived an error upper bound using Lax-Milgram theorem for both weak 4DVAR and 3DVAR constraint weak 4DVAR. We analysed the results for various parameter values. Then we verified these results with numerical experiments for linear and Lorenz'63 system for various parameter regimes. Although the results presented are simplistic however, their concurrence with the analytical results promises the possibility of their extension to larger, more complex systems.

6.2 Future Directions

In this section we discuss few directions in which analysis of data assimilation schemes can be built upon. All through out this thesis we only considered low dimensional models to test our proposals. Applying data assimilation algorithms to large scale models present numerous computational challenges. A natural direction for this work to evolve would be to investigate, how the method performs when applied to more complex, high dimensional problems.

Another direction would be to consider more realistic parameters for the filtering scheme such that nonlinear observation operator or correlations between observed variables. Similarly the background noise was considered uncorrelated however in a realistic settings that assumption might not hold so investigating such cases is another direction in which this work can be extended.

More specifically the analytical discussions in Chapters 2 and 3 have focused upon 3DVAR filtering scheme to show the convergence of the estimate to the true underlying state . Extending this analysis to more complex filtering schemes, such as ExKF, EnKF etc. is a possibility. Similarly a study of more complex underlying systems is another desirable direction for future work.

Finally the variational schemes proposed in Chapters 4 and 5 require a more sophisticated

and structured approach for bias calculation and derivation of error upper bounds for complex non-linear models.

Bibliography

[1] H. Abarbanel. *Predicting the Future: Completing Models of Observed Complex Systems*. Springer. Series: Understanding Complex Systems, 2013.

[2] S. Akella and I. M. Navon. Different approaches to model error formulation in 4d-var: a study with high-resolution advection schemes. *Tellus A*, 61(1):112–128, 2009.

[3] A. Apte, C. Jones, A. Stuart, and J. Voss. Data assimilation: Mathematical and statistical perspectives. *International Journal for Numerical Methods in Fluids*, 56(8):1033–1046, 2008.

[4] A. Azouani, E. Olson, and E. Titi. Continuous data assimilation using general interpolant observables. *Journal of Nonlinear Science*, 24:277–304, 2014.

[5] G. Benettin, L. Galgani, and J. Strelcyn. Kolmogorov entropy and numerical experiments. *Phys. Rev. A*, 14:2338–2345, Dec 1976.

[6] A. Bennett. *Inverse Modeling of the Ocean and Atmosphere*. Cambridge University Press, 2003.

[7] K. Bergemann and S. Reich. An ensemble kalman-bucy filter for continuous data assimilation. *Meteorologische Zeitschrift*, 21(3):213–219, 2012.

[8] D. Bloemker, K. Law, A. Stuart, and K. Zygalakis. Accuracy and stability of the continuous-time 3DVAR filter for the navier-stokes equation. *Nonlinearity*, 2014.

[9] M. Bocquet and P. Sakov. Combining inflation-free and iterative ensemble kalman filters for strongly nonlinear systems. *Nonlin. Processes Geophys.*, 19:383–399, 2012.

[10] F. Bouttier and P. Courtier. Data assimilation concepts and methods. Meteorological Training Course Lecture Series, 1999.

[11] C. Brett, K. Lam, K. Law, D. McCormick, M. Scott, and A. Stuart. Accuracy and stability of filters for dissipative pdes. *PhysicaD: Nonlinear Phenomena*, 2013.

[12] P. Courtier, E. Andersson, W. Heckley, D. Vasiljevic, M. Hamrud, A. Hollingsworth, F. Rabier, M. Fisher, and J. Pailleux. The ecmwf implementation of three-dimensional variational

assimilation (3d-var). i: Formulation. *Quarterly Journal of the Royal Meteorological Society*, 124(550):1783–1807, 1998.

[13] M. Cullen, M. Freitag, and S. Kindermann. *Large Scale Inverse Problems. Computational Methods and Applications in the Earth Sciences*. De Gruyter, 2013.

[14] D. P. Dee and A. M. Da Silva. Data assimilation in the presence of forecast bias. *Quarterly Journal of the Royal Meteorological Society*, 124(545):269–295, 1998.

[15] J. C. Derber. A variational continuous assimilation technique. *Monthly weather review*, 117(11):2437–2446, 1989.

[16] A. Doucet, N. De Freitas, and N. Gordon. *Sequential Monte Carlo methods in practice*. Springer Verlag, 2001.

[17] G. Evensen. Using the extended kalman filter with a multilayer quasi-geostrophic ocean model. *Journal of Geophysical Research: Oceans*, 97(C11):17905–17924, 1992.

[18] G. Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99(C5):10143–10162, 1994.

[19] G. Evensen. The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean dynamics*, 53(4):343–367, 2003.

[20] G. Evensen. *Data Assimilation: the Ensemble Kalman Filter*. Springer Verlag, 2009.

[21] M. Fisher. Generalized frames on the sphere, with application to the background-error covariance modelling. In *Proc. ECMWF Seminar on Recent Developments in Numerical Methods for Atmospheric and Ocean Modelling*, pages 87–102, 2004.

[22] C. Foias, M. Jolly, L. Kukavica, and E. Titi. The lorenz equationas a metaphor for the navier-stokes equation, discrete and continous dynamical systems. *Discrete and Continous Dynamical Systems*, 7(4):403–429, 2001.

[23] C. Foias and G. Prodi. Sur le comportement global des solutions nonstationnaires des équations de navier-stokes en dimension 2. *Rend. Sem. Mat. Univ. Padova*, 39(1), 1967.

[24] R. Furrer and T. Bengtsson. Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants. *Journal of Multivariate Analysis*, 98(2):227–255, 2007.

[25] M. Goodliff, J. Amezcua, and P. J. V. Leeuwen. Comparing hybrid data assimilation methods on the lorenz 1963 model with increasing non-linearity. *Tellus A*, 67(0), 2015.

[26] A. K. Griffith and N. K. Nichols. Adjoint methods in data assimilation for estimating model error. *Flow, turbulence and combustion*, 65(3-4):469–488, 2000.

[27] M. Hairer, A. Stuart, and J. Voss. Sampling conditioned diffusions. In J. Blath, P. Mörters, and M. Sceutzow, editors, *Trends in Stochastic Analysis*, LMS Lecture Notes 353. Cambridge University Press, 2008.

[28] A. Harvey. *Forecasting, Structural Time Series Models and the Kalman filter*. Cambridge Univ Pr, 1991.

[29] K. Hayden, E. Olson, and E. Titi. Discrete data assimilation in the Lorenz and 2d Navier-Stokes equations. *Physica D: Nonlinear Phenomena*, pages 1416–1425, 2011.

[30] P. L. Houtekamer and H. L. Mitchell. Data assimilation using an ensemble kalman filter technique. *Monthly Weather Review*, 126(3):796–811, 1998.

[31] B. R. Hunt, E. Kalnay, E. J. Kostelich, E. Ott, D. J. Patil, T. Sauer, I. Szunyogh, J. A. Yorke, and A. V. Zimin. Four-dimensional ensemble kalman filtering. *Tellus A*, 56(4):273–277, 2004.

[32] A. Jazwinski. *Stochastic processes and filtering theory*. Academic Pr, 1970.

[33] C. Johnson, N. K. Nichols, and B. J. Hoskins. Very large inverse problems in atmosphere and ocean modelling. *International Journal for Numerical Methods in Fluids*, 47(8-9):759–771, 2005.

[34] D. Jones and E. Titi. Upper bounds on the number of determining modes, nodes, and volume elements for the navier-stokes equations. *Indiana University Mathematics Journal*, 42(3):875–888, 1993.

[35] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME - Journal of Basic Engineering*, (82 (Series D)):35–45, 1960.

[36] E. Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 2003.

[37] E. Kalnay, H. Li, T. Miyoshi, S. Yang, and J. Ballabrera-Poy. 4-d-var or ensemble kalman filter? *Tellus A*, 59(5):758–773, 2007.

[38] M. Kostuk. *Synchronization and statistical methods for the data assimilation of HVc neuron models*. PhD thesis, University of California, San Diego, 2012.

[39] K. Law, D. Sanz-Alonso, A. Shukla, and A. Stuart. Filter accuracy for the lorenz 96 model: fixed versus adaptive observation operators. *Physica D: Nonlinear Phenomena*, 325:1–13, 2016.

[40] K. Law, A. Shukla, and A. Stuart. Analysis of the 3dvar filter for the partially observed lorenz '63 model. *Discrete and Continuous Dynamical Systems A*, 34:1061–1078, 2014.

[41] K. Law, A. Shukla, and A. Stuart. Analysis of the 3dvar filter for the partially observed lorenz '63 model. *Discrete and Continuous Dynamical Systems A*, 34:1061–1078, 2014.

[42] K. Law and A. Stuart. Evaluating data assimilation algorithms. *Monthly Weather Review*, 2012.

[43] K. Law, A. Stuart, and K. Zygalakis. *Data Assimilation: A Mathematical Introduction*. Lecture Notes, 2014.

[44] F.-X. Le Dimet and V. Shutyaev. On deterministic error analysis in variational data assimilation. *Nonlinear Processes in Geophysics*, 12(4):481–490, 2005.

[45] F. X. Le Dimet and O. Talagrand. Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A*, 38A(2):97–110, 1986.

[46] B. Lemieux and A. Vidard. Implementation of the weak constraint 4D-Var in NEMOVAR. Contract D3.2.1 & D3.2.2, 2012.

[47] J. Lewis, S. Lakshmivarahan, and S. Dhall. *Dynamic Data Assimilation: A Least Squares Approach*. Number v. 13 in Dynamic data assimilation: a least squares approach. Cambridge University Press, 2006.

[48] J. M. Lewis and J. C. Derber. The use of adjoint equations to solve a variational adjustment problem with advective constraints. *Tellus A*, 37(4):309–322, 1985.

[49] C. Liu, Q. Xiao, and B. Wang. An ensemble-based four-dimensional variational data assimilation scheme. part i: Technical formulation and preliminary test. *Monthly Weather Review*, 136(9):3363–3373, 2008.

[50] A. C. Lorenc. Analysis methods for numerical weather prediction. *Quart. J. R. Met. Soc.*, 112(474):1177–1194, 2000.

[51] A. C. Lorenc. Modelling of error covariances by 4d-var data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 129(595):3167–3182, 2003.

[52] A. C. Lorenc, S. P. Ballard, R. S. Bell, N. B. Ingleby, P. L. F. Andrews, D. M. Barker, J. R. Bray, A. M. Clayton, T. Dalby, D. Li, T. J. Payne, and F. W. Saunders. The Met. Office global three-dimensional variational data assimilation scheme. *Quart. J. R. Met. Soc.*, 126(570):2991–3012, 2000.

[53] E. Lorenz. Deterministic nonperiodic flow. *Atmos J Sci*, 20:130–141, 1963.

[54] E. Lorenz. Predictability: A problem partly solved. In *Proc. Seminar on Predictability*, volume 1, pages 1–18, 1996.

[55] E. Lorenz and K. Emanuel. Optimal sites for supplementary weather observations: Simulation with a small model. *Journal of the Atmospheric Sciences*, 55:399–414, 1998.

[56] A. Majda and J. Harlim. *Filtering Complex Turbulent Systems.* Cambridge University Press, 2012.

[57] A. Majda, J. Harlim, and B. Gershgorin. Mathematical strategies for filtering turbulent dynamical systems. *Dynamical Systems*, 27(2):441–486, 2010.

[58] X. Mao. *Stochastic Differential Equations And Applications.* Horwood, 1997.

[59] D. Oliver, A. Reynolds, and N. Liu. *Inverse Theory for Petroleum Reservoir Characterization and History Matching.* Cambridge University Press, 2008.

[60] E. Olson and E. Titi. Determining modes for continuous data assimilation in 2D turbulence. *Journal of statistical physics*, 113(5):799–840, 2003.

[61] E. Ott, B. Hunt, I. Szunyogh, A. Zimin, E. Kostelich, M. Corazza, E. Kalnay, D. Patil, and J. Yorke. A local ensemble kalman filter for atmospheric data assimilation. *Tellus A*, 56(5):415–428, 2004.

[62] D. F. Parrish and J. C. Derber. The national meteorological center's spectral statistical-interpolation analysis system. *Monthly Weather Review*, 120(8):1747–1763, 1992.

[63] L. Pecora and T. Carroll. Synchronization in chaotic systems. *Physical review letters*, 64(8):821–824, 1990.

[64] C. Pires, R. Vautard, and O. Talagrand. On extending the limits of variational assimilation in nonlinear chaotic systems. *Tellus A*, 48(1), 2011.

[65] J. Quinn and H. Abarbanel. State and parameter estimation using monte carlo evaluation of path integrals. *Quarterly Journal of the Royal Meteorological Society*, 2010.

[66] F. Rabier. Overview of global data assimilation developments in numerical weather-prediction centres. *Quarterly Journal of the Royal Meteorological Society*, 131(613):3215–3233, 2005.

[67] F. Rabier, A. McNally, E. Andersson, P. Courtier, P. Undén, J. Eyre, A. Hollingsworth, and F. Bouttier. The ecmwf implementation of three-dimensional variational assimilation (3d-var). ii: Structure functions. *Quarterly Journal of the Royal Meteorological Society*, 124(550):1809–1829, 1998.

[68] J. C. Robinson. *Infinite-Dimensional Dynamical Systems.* Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 2001.

[69] Y. Sasaki. An objective analysis based on the variational method. *Journal of the Meteorological Society of Japan. Ser. II*, 36(3):77–88, 1958.

[70] C. Sparrow. *The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors.* Springer, 1982.

[71] A. M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numer.*, 19:451–559, 2010.

[72] K. Swanson, T. Palmer, and R. Vautard. Observational error structures and the value of advanced assimilation techniques. *Journal of the atmospheric sciences*, 57(9):1327–1340, 2000.

[73] T. Tarn and Y. Rasis. Observers for nonlinear stochastic systems. *Automatic Control, IEEE Transactions*, 21(4):441–488, 1976.

[74] R. Temam. *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, volume 68 of *Applied Mathematical Sciences*. Springer-Verlag, New York, second edition, 1997.

[75] Y. Trémolet. Accounting for an imperfect model in 4d-var. *Quarterly Journal of the Royal Meteorological Society*, 132(621):2483–2504, 2006.

[76] Y. Trémolet. Model-error estimation in 4d-var. *Quarterly Journal of the Royal Meteorological Society*, 133(626):1267–1280, 2007.

[77] A. Trevisan, M. D'Isidoro, and O. Talagrand. Four-dimensional variational assimilation in the unstable subspace and the optimal subspace dimension. *Quarterly Journal of the Royal Meteorological Society*, 136(647):487–496, 2010.

[78] A. Trevisan and F. Uboldi. Assimilation of standard and targeted observations within the unstable subspace of the observation analysis forecast cycle system. *Journal of the Atmospheric Sciences*, 61(1):103–113, 2004.

[79] W. Tucker. A rigorous ode solver and smale's 14th problem. *Journal of Foundations of Computational Mathematics*, 2:53–117, 2002.

[80] P. Van Leeuwen. Particle filtering in geophysical systems. *Monthly Weather Review*, 137:4089–4114, 2009.

[81] C. K. Wikle and L. M. Berliner. A bayesian tutorial for data assimilation. *Physica D: Nonlinear Phenomena*, 230(1):1–16, 2007.

[82] C. K. Williams. Prediction with gaussian processes: From linear regression to linear prediction and beyond. In *Learning in graphical models*, pages 599–621. Springer, 1998.

[83] W. Yang, I. M. Navon, and P. Courtier. A new hessian preconditioning method applied to variational data assimilation experiments using nasa general circulation models. *Monthly Weather Review*, 124(5):1000–1017, 1996.

[84] W.-C. Yueh. Eigenvalues of several tridiagonal matrices. *Applied Mathematics E-Notes*, 5(66-74):210–230, 2005.